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

Benjia Zhou[1,2*], Pichao Wang[2†], Jun Wan[1,3,4†], Yanyan Liang[1], Fan Wang[2],
Du Zhang[1], Zhen Lei[3,4], Hao Li[2], Rong Jin[2]
[1]Macau University of Science and Technology [2]Alibaba Group
[3]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

Abstract

Decoupling spatiotemporal representation refers to decomposing the spatial and temporal features into dimension-independent factors. Although previous RGB-D-based motion recognition methods have achieved promising performance through the tightly coupled multi-modal spatiotemporal representation, they still suffer from (i) optimization difficulty under small data setting due to the tightly spatiotemporal-entangled modeling; (ii) information redundancy as it usually contains lots of marginal information that is weakly relevant to classification; and (iii) low interaction between multi-modal spatiotemporal information caused by insufficient late fusion. To alleviate these drawbacks, we propose to decouple and recouple spatiotemporal representation for RGB-D-based motion recognition. Specifically, we disentangle the task of learning spatiotemporal representation into 3 sub-tasks: (1) Learning high-quality and dimension independent features through a decoupled spatial and temporal modeling network. (2) Recoupling the decoupled representation to establish stronger space-time dependency. (3) Introducing a Cross-modal Adaptive Posterior Fusion (CAPF) mechanism to capture cross-modal spatiotemporal information from RGB-D data. Seamless combination of these novel designs forms a robust spatiotemporal representation and achieves better performance than state-of-the-art methods on four public motion datasets. Our code is available at https://github.com/damo-cv/MotionRGBD.

1. Introduction

The RGB-D-based motion recognition has attracted much attention in computer vision due to its broad application scenarios such as video surveillance and human-object interfaces. Recently, the CNN and RNN based methods greatly improve the performance of recognition on both gesture [1, 27, 37, 43, 45] and action [7, 17, 38, 41] through fully exploring the color and depth cues. Meanwhile, inspired by the transformer scaling success in vision tasks, Transformer-based methods [10, 21] also achieve surprising results on RGB-D-based motion recognition by introducing the cross-attention module for multi-modality fusion.

Although these works make great progress, we find they are still problematic in the following three aspects. (i) **Optimization difficulty** exists in the case of limited RGB-D data due to the tightly spatiotmporal entangled modelling (*e.g.*, C3D [33] and I3D [3]). (ii) **Redundant information** is hard to deal with in the entangled space-time space. To address the above two issues, some decoupled networks (*i.e.*, 2D CNN+LSTM/Transformer [19, 34]) are proposed to learn the spatiotemporal independent representation. However, we argue that these methods are not conducive to compact representation as they somewhat weaken or even de-

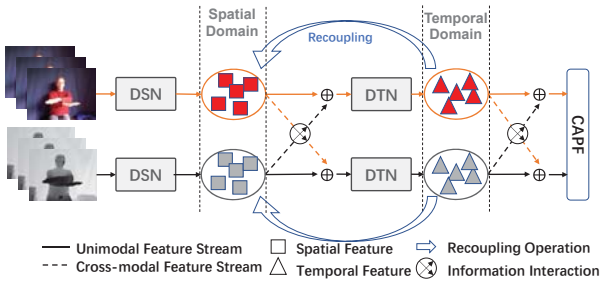

Figure 1. Illustration of the proposed multi-modal spatiotemporal representation learning framework. The RGB-D-based motion recognition can be described as cross-modal representation interactive learning based on decoupled and recoupled spatiotemporal information. Wherein DSN and DTN represent **d**ecoupled **s**patial and **d**ecoupled **t**emporal feature learning **n**etworks, respectively; And $\oplus$ represents the element-wise add operation.

*Work done during an internship at Alibaba Group.
†Corresponding author.

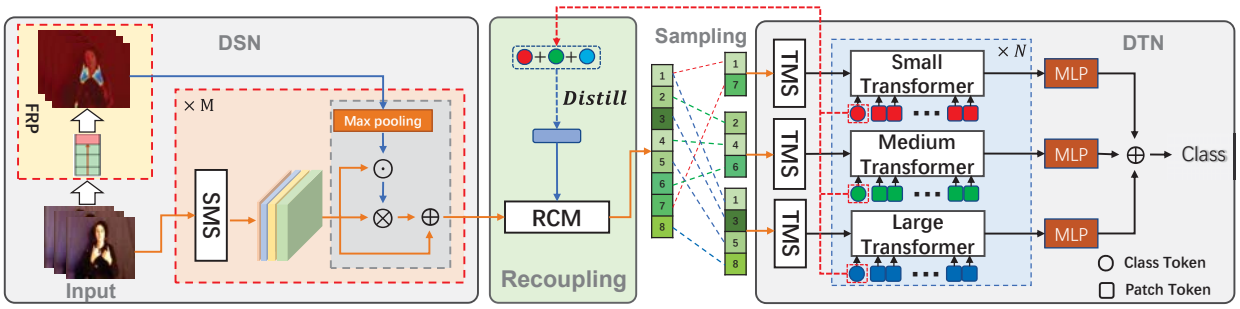

Figure 2. Illustration of proposed decoupling and recoupling saptiotemporal representation learning network. The whole network mainly consists of a decoupled spatial and temporal representation learning networks namely DSN and DTN, as well as a spatiotemporal recoupling module (RCM). The FRP indicates a fast regional positioning module; SMS and TMS indicate the space- and time-centric multi-scale Inception Module respectively. $\odot$, $\otimes$ and $\oplus$ indicate element-wise product, 1D convolution and element-wise add operation respectively.

stroy the original spatiotemporal coupling structure. Considering that a certain number of human action classes have strong correlations between time and space, the recoupling process after spatiotemporal decoupling is still necessary. (iii) **Insufficient interaction** occurs between multi-modal spatiotemporal information. Several works [45, 47] adopt independent branches for unimodal spatiotemporal representations learning followed by late fusion, resulting in insufficient cross-modal information communication. Thus, it is still a challenge to explore high quality multi-modal spatiotemporal features.

Given the aforementioned concerns, as illustrated in Figure 1, we introduce a new method of multi-modal spatiotemporal representation learning for RGB-D-based motion recognition. It mainly consists of a decoupled spatial representation learning network (DSN), a decoupled temporal representation learning network (DTN) and a cross-modal adaptive posterior fusion module (CAPF). For each unimodal branch, as shown in Figure 2, we propose a decoupling and recoupling spatiotemporal feature learning method, wherein a spatiotemporal recoupling module (RCM) is designed as a bonding of DSN and DTN. RCM acts as feature selection for DSN and knowledge integration for DTN. The entire framework can be decomposed into 3 steps: (1) **Spatiotemporal decoupling learning**. In the DSN, the video clips are first fed into stack of inception-based spatial multi-scale features learning (SMS) modules to extract hierarchical spatial features. Meanwhile, they are also input into a bypath network, called fast regional positioning module (FRP), to generate visual guidance map, which guides the network to focus on local important areas in the video frame. Then the integrated spatial features from SMS and FRP are fed into RCM for feature selection. After that, we sample several sub-sequences at different frame rates from the enhanced spatial features as input to the DTN. The DTN is configured as a multi-branch structure with an inception-based temporal multi-scale layer (TMS) and multiple Transformer blocks for hierarchical local fine-grained and global coarse-grained temporal feature learning. (2) **Spatiotemporal recoupling learning**. To rebuild the space-time interdependence, a self-distillation-based recoupling strategy is developed. As shown by the red dashed line in Figure 2, the recoupling method is designed as an inner loop optimization mechanism to distill the inter-frame correlations from time domain into the space domain, to enhance the quality of the spatial features via RCM. (3) **Cross-modal interactive learning**. For multi-modal representation learning from RGB-D data, as shown in Figure 1, we propose an interactive cross-modal spatiotemporal representation learning method. Specifically, the cross-modal spatial features derived from unimodal branches firstly interact at the spatial level and are mapped to a joint spatial representation. Then it is separately integrated with the two unimodal spatial feature streams through the residual structure. After that, the two spatial feature streams are input into their respective temporal modeling networks to capture temporal features. Similar to the spatial feature interaction, a joint temporal representation can also be obtained through interaction at the temporal level. Combined with the joint temporal representation, the two temporal feature streams are fed into the CAPF, which is based on a multi-loss joint optimization mechanism, to conduct deep multi-modal representation fusion.

Through the above design, our method not only effectively achieves the spatiotemporal information decoupling and recoupling learning within each modality, but also realizes the deep communication and fusion of multi-modal spatiotemporal information. The proposed method achieves state-of-the-art performance on four public RGB+D gesture/action datasets, namely NvGesture [26], Chalearn

IsoGD [36], THU-READ [32], and NTU-RGBD [28].

2. Related Work

2.1. Motion Recognition based on RGB-D Data

Recently, with the availability of low-cost RGB-D sensors, RGB-D-based motion recognition has attracted extensive attention. To effectively encode the robust multi-modal spatiotemporal information for motion recognition, Zhu *et al*. [46] presents a pyramidal-like 3D convolutional network structure for spatiotemporal representation extraction and fusion. Kong *et al*. [20] propose to compress and project the RGB-D data into a shared space to learn cross-modal features for effective action recognition. Yu *et al*. [43] employ NAS to search for modal-related network structures and optimal multi-modal information transmission path for RGB-D data. Different from the modal-separated multi-branch networks, scene flow is adopted in [39] for compact RGB-D representation learning. Wang *et al*. [41] propose to use a single network c-ConvNet for multi-modal spatiotemporal representation learning and aggregation. Unlike previous methods that interact with the cross-modal information on coupled spatiotemporal information, we focus on the interaction of multi-modal features in two independent dimensions of space and time.

2.2. Decoupled Spatiotemporal Feature Learning

Considering the importance of decoupled spatiotemporal feature learning in sequence, Shi *et al*. [31] present a decoupled spatiotemporal attention network (DSTA-Net) for skeleton-based action recognition. Liu *et al*. [22] present a decoupled spatiotemporal Transformer (DSTT) architecture to improve video inpainting tasks. He *et al*. [14] propose an effective spatiotemporal network StNet, which employs separated channel-wise and temporal-wise convolution operations for decoupled local and global representation learning. Zhang *et al*. [44] present a hierarchically decoupled spatiotemporal contrastive learning method for self-supervised video representation learning. They capture the spatial and temporal features by decoupling the learning objective into two contrastive sub-tasks, and perform it hierarchically to encourage multi-scale understanding. In contrast, in this work, we target to learn recoupled features due to strong correlations between time and space for some human actions. Thus, a distillation-based recoupling process is introduced on the decoupled spatiotemporal features.

3. Proposed Method

In this paper, we assume the spatiotemporal representation can be decomposed into two sub-domains: the spatial domain that correlates with the visual information, and the temporal domain that describes the time-related concept. Based on this, for unimodal spatiotemporal representation learning, we first decouple the spatiotemporal modeling process to learn domain-independent representations (Sec. 3.1). Then a recoupling method is introduced based on the inner loop optimization mechanism during the training stage (Sec. 3.2), to strengthen the spatiotemporal connection. For multi-modal features interactive learning, we first separately integrate the cross-modal spatiotemporal information into the spatial and temporal domains, and then employ an adaptive posterior fusion mechanism to further fuse the multi-modal features (Sec.3.3).

3.1. Decoupling Spatiotemporal Representation

3.1.1 Decoupled Spatial Feature Learning (DSN)

As shown in Figure 2, the DSN is composed of fast regional positioning module (FRP) and stack of inception-based spatial multi-scale features learning (SMS) modules. Let $[I_1, I_2, \ldots, I_T]$ denote the input with length T sampled from the video. It is fed into the SMS and FRP modules in parallel to capture the hierarchical spatial features and generate visual guidance maps.

SMS Module. The SMS module is composed of a space-centric 3D Inception Module[1] [3] and a Max Pooling layer. It extracts the multi-scale features of m-th frame at layer l-th by:

$$f_m^l = \begin{cases} \text{Maxpool}(\mathcal{C}_{S-Inc}(I_m, W)), & l = 1 \\ \text{Maxpool}(\mathcal{C}_{S-Inc}(f_m^{l-1}, W)), & l > 1 \end{cases} \quad (1)$$

where f_m^{l-1} represents the output of the previous layer; $\mathcal{C}_{S-Inc}(\cdot, W)$ indicates the Inception Module with learnable parameter matrix W; Maxpool is the Max Pooling layer. Meanwhile, to guide each SMS module to focus on local important areas in the image, a visual guidance map is embedded in parallel to it to further enhance its visual perception, which is the proposed FRP module.

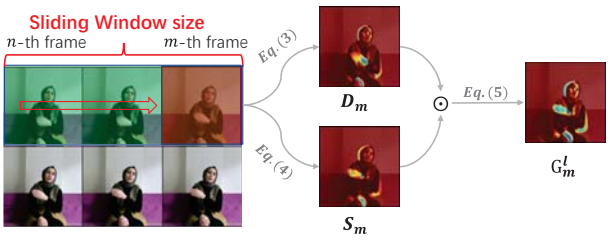

Figure 3. Overview of the proposed fast regional positioning module FRP. In each sliding window, the FRP module locates important areas in the last frame according to the successive video frames before it.

FRP Module. The FRP module is designed to generate the visual guidance map to rapidly locate important areas

[1] The size of the convolution kernel in the temporal dimension is 1.

in an image for motion recognition. The visual guidance map is obtained directly by combining static and dynamic guidance maps. Specifically, as shown in Figure 3, within a sliding window from the n-th to the m-th frame, we first calculate the dynamic image $DI(n,m)$ [45]:

$$\begin{aligned} DI(n,m) =& DI(n-1,m-1) + (m-n)\times \\ & (I_{n-1}+I_m) - 2\sum_{l=n}^{m-1} I_l, \quad s.t. \quad m>n \end{aligned} \tag{2}$$

where $DI(n-1,m-1)$ represents the dynamic image from the previous sliding window. Then the dynamic guidance map on m-th frame can be obtained by:

$$D_m = \delta(DI(n,m)) \times \lambda \tag{3}$$

where δ and λ indicate activation function and signal amplification factor respectively, herein we use GELU [15] and $\lambda = 2$ in all of experiments. However, we find that D_m is sensitive to lighting as it only considers the motion information between multiple frames. To mitigate this problem, we introduce the static guidance map S_m, which can be defined as:

$$\begin{aligned} S_m &= \text{dilate}(\text{erode}(\hat{D}_\text{m}), (k\times k))) \\ \hat{D}_\text{m} &= \begin{cases} D_{\text{m,(i,j)}} & D_{\text{m,(i,j)}} >= D_{\text{m,mean}} \\ 0 & otherwise \end{cases} \end{aligned} \tag{4}$$

where dilate(erode$(\cdot,(k\times k)))$ indicates the dilation and erosion operations with a kernel size of $k\times k$ in binary mathematical morphology. $\hat{D}_\text{m}$ is an attention matrix, in which all elements except for these higher than the mean value are set to zero, as we empirically observe that the response value at the region affected more by lighting is generally below average. After that, combining the Eq.3 and 4, the visual guidance map of m-th frame at the l-th layer of the network can be derived by:

$$G_m^l = \text{Maxpool}((D_m + S_m)\times S_m) \tag{5}$$

where the Max Pooling operation is used to scale the size of the current guidance map G_m^l to match the feature map f_m^l. In addition, we also perform the normalization and alignment operations for the guidance map G_m^l, which have been discussed in detail in the supplementary material. Finally, the spatial features $O \in \mathbb{R}^{T\times d}$ with length T and dimension d output from the DSN can be simply formulated as:

$$\begin{aligned} & O = [O_1^l, O_2^l, \ldots, O_T^l], \quad \forall l = 1,2,\ldots,M \\ & s.t. \quad O_m^l = (f_m^l \odot G_m^l) \otimes f_m^l + f_m^l \end{aligned} \tag{6}$$

where M represents the number of total network layers used in DSN; $\odot$ and $\otimes$ represent element-wise product and 1D convolution operations respectively; And O_m^l means the result of fusing the raw feature stream f_m^l with the visual guidance map G_m^l. It is then enhanced via RCM module for decoupled temporal representation learning.

3.1.2 Decoupled Temporal Feature Learning (DTN)

The temporal representation learning network DTN takes the **enhanced spatial features** $\hat{O} \in \mathbb{R}^{T\times d/2}$ (formal definition in Sec.3.2) as the input. As shown in Figure 2, the DTN is configured as a multi-branch and two-stage structure to progressively learn the hierarchical temporal representation at local fine-grained level and global coarse-grained level. Specifically, for a single sub-branch k, we first sample a sub-sequence of features $\hat{O}_k \in \mathbb{R}^{T_n\times d/2}$ of length T_n from $\hat{O}$ by a discrete sampling strategy:

$$\hat{O}_k = \{\hat{O}_\tau | \tau = \mathcal{R}[\lceil\frac{T}{T_n}\rceil\times t-1, \lceil\frac{T}{T_n}\rceil\times t], t=1,2,\ldots,T_n\} \tag{7}$$

where the $\mathcal{R}[a,b]$ represents randomly selecting an integer $x, s.t.\ a\leqslant x\leqslant b$; $\hat{O}_k$ serves as the input of temporal multi-scale features learning module (TMS).

TMS Module. The TMS is composed of a time-centric 3D Inception Module[2] $\mathcal{C}_{T-Inc}(\cdot, W_k)$ and Max Pooling layer. We only use one layer TMS to capture the local fine-grained spatial features $\hat{O}_k^L$:

$$\hat{O}_k^L = \text{Maxpool}(\mathcal{C}_{T-Inc}(\hat{O}_k, W_k)) \tag{8}$$

After that, $\hat{O}_k^L$ is fed into stack of Transformer blocks to learn the coarse-grained temporal representation.

Transformer block. To reduce the redundant marginal information in captured temporal features, we utilize a Transformer structure based on k-NN multi-head self-attention layer [42]. Thus the modeling process of each Transformer block can be formulated as:

$$\hat{O}_k^G = \text{MLP}(\text{LN}(\text{MSA}_{kNN}(\hat{O}_k^L))) + \hat{O}_k^{G-1} \tag{9}$$

where $\hat{O}_k^{G-1}$ indicates the output feature of the previous layer; $\text{MSA}_{kNN}(\cdot)$ indicates the k-NN multi-head self-attention layer and LN represents layer normalization. Furthermore, to avoid overfitting to one of the sub-branches, we introduce a temperature parameter τ to control the sharpness of the output distribution of each sub-branch and impose a constraint loss on it. Therefore, the output of the DTN network can be formulated as:

$$O^{CLS} = \sum_{k=1}^{K} \text{MLP}(O_k^{CLS})/\tau, \forall k = 1,2,3,\ldots,K \tag{10}$$

where O^{CLS} is the class token vector embedded in the Transformer block and K indicates the number of sub-branches. τ follows a cosine schedule from 0.04 to 0.07 during the training. We analytically demonstrate in Sec.5.1 that the tactics of k-NN Attention, sharpness and multi-loss can bring performance gains for the proposed network.

[2]The size of the convolution kernel in the spatial dimension is 1×1.

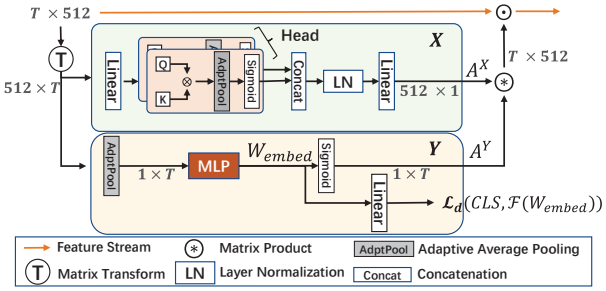

Figure 4. Overview of the proposed recoupling module RCM.

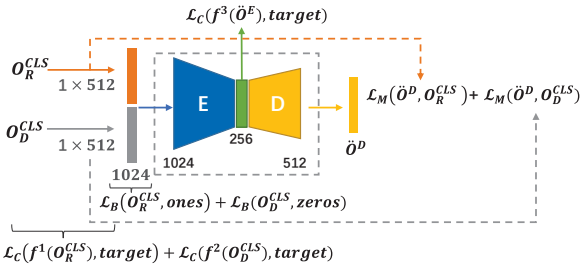

Figure 5. Adaptive fusion module CAPF. $\mathcal{L}_C$, $\mathcal{L}_B$, $\mathcal{L}_M$ indicate cross entropy, binary cross entropy and mean square error loss functions, respectively; $traget$ represents ground-truth; E and D indicate the Encoder and Decoder respectively.

Based on spatial features O from DSN and temporal features O^{CLS} from DTN, we propose a recoupling strategy to strengthen the space-time connection during training, because the spatiotemporal decoupled learning method weakens the original coupled structure.

3.2. Recoupling Spatiotemporal Representation

The recoupling strategy is developed to rebuild the space-time interdependence, which reversely applies the distilled inter-frame correlations from time domain into the space domain through a self-distillation-based inner loop optimization mechanism during the training. Specifically, the spatial feature stream $O \in \mathbb{R}^{T\times d}$ derived from the spatial features learning network DSN is first linearly mapped into a low-dimensional space. Then the mapped features $\bar{O} \in \mathbb{R}^{T\times d/2}$ is transposed and fed into the recoupling module RCM.

RCM module. The RCM module is configured as dual pathway to enhance the spatial features from the X (feature dimension) and Y (sequence dimension) directions, as shown in Figure 4. For X direction (intra-frame), inspired by *self-attention* [8], we utilize a set of learnable matrices: $W_Q \in \mathbb{R}^{T\times T}$ and $W_K \in \mathbb{R}^{T\times T}$ to calculate the attention map A_X (Note that if multiple heads are configured, they will be concatenated and mapped to a fixed dimension).

$$\begin{aligned} &Q = \bar{O}^T W_Q, \quad K = \bar{O}^T W_K, \\ &A_X = \delta(\mathrm{GAP}(\frac{QK^T}{\sqrt{d}})), \quad A_X \in \mathbb{R}^{1\times d/2} \end{aligned} \tag{11}$$

where Q and K denote the queries and keys, δ is the Sigmoid activation function, and GAP indicates the adaptive global average pooling operation. The attention map A_X describes the correlation of the intra-frame in the sequence. For Y direction (inter-frame), the $\bar{O}$ is first compressed along the channel dimension through GAP operation to obtain a feature vector with a dimension of $1 \times T$. Then it passes through an MLP block with two hidden layers to obtain a weight embedding W_{embed}.

$$W_{embed} = \mathrm{MLP}(\mathrm{GAP}(\bar{O}^T)), \quad W_{embed} \in \mathbb{R}^{1\times T} \tag{12}$$

Finally, it is combined with a Sigmoid activation function to describe the correlation of the inter-frame in the sequence.

$$A_Y = \delta(W_{embed}), \quad A_Y \in \mathbb{R}^{1\times T} \tag{13}$$

However, the MLP layer cannot effectively learn the inter-frame correlation from the captured spatial features $\bar{O}$. Therefore, we employ a self-distillation loss function $\mathcal{C}_d$ to introduce additional supervision for the MLP block to distill inter-sequence correlation knowledge from temporal domain into W_{embed}. The distillation process can be formulated as:

$$\begin{aligned} &\mathcal{C}_d = \mathcal{L}_d(CLS, \mathcal{F}(W_{embed})), CLS = \sum_{k=1}^{K} O_k^{CLS} \\ &s.t. \quad \mathcal{L}_d(x, y) = \frac{1}{N_B}\sum_{i=1}^{N_B} KL(x_i/\mathcal{T} - y_i/\mathcal{T}) \end{aligned} \tag{14}$$

where $\mathcal{T}$ is the distillation temperature parameter, N_B is the batch size, KL indicates Kullback-Leibler divergence [16] and $\mathcal{F}$ indicates linear mapping function. Combining Eq.11 and 13, the attention map for spatial feature enhancement can be derived as:

$$A_{XY} = \sum_{i=1}^{I}\sum_{j=1}^{J} A_{X,i}^T \times A_{Y,j}, A_{XY} \in \mathbb{R}^{T\times d/2} \tag{15}$$

where I and J indicate the element index in A_X and A_Y respectively. It is then applied to the raw spatial feature stream O by the element-wise product operation:

$$\hat{O} = O \odot A_{XY} \tag{16}$$

where $\hat{O}$ represents the enhanced spatial feature, which is used for temporal representation learning.

3.3. Cross-modal Interactive Learning

For the RGB-D multi-modal spatiotemporal features separately extracted from the two network branches, we

propose to interact them at the spatial and temporal dimensions respectively at first, as shown in Figure 1. Take the RGB modality as an example, the two spatial features from the RGB-D modalities are first interacted through the MLP layer to generate a joint spatial representation. It is then integrated with raw spatial features of the RGB modality by a residual structure:

$$\ddot{O}_R^S = \text{LN}(\text{MLP}([\hat{O}_R^S||\hat{O}_D^S])) + O_R^S \quad (17)$$

where $\hat{O}_R^S$ and $\hat{O}_D^S$ denote the enhanced spatial features of RGB-D modalities respectively; And $||$ represents concatenation. Similar to the interactive way of spatial features, after obtaining a joint temporal representation $\ddot{O}^T \in \mathbb{R}^{1\times d/2}$, it is fed into the cross-modal adaptive posterior fusion (CAPF) module for deep multi-modal features fusion.

CAPF module. As shown in Figure 5, the CAPF module contains an Encoder and Decoder, which are composed of MLP blocks with multiple hidden layers. The $\ddot{O}^T$ is first fed into the Encoder to generate the encoded embedding $\ddot{O}^E \in \mathbb{R}^{1\times d/4}$ which can be used for classification. Then $\ddot{O}^E$ passes through the Decoder to obtain a decoded embedding $\ddot{O}^D \in \mathbb{R}^{1\times d/2}$ which can be used to supervise the Encoder. Therefore, the final classification score of the entire network can be obtained by:

$$C = \text{Argmax}(f^1(O_R^{CLS}) + f^2(O_D^{CLS}) + f^3(\ddot{O}^E)) \quad (18)$$

where f^* denotes the linear classifier and Argmax is used to take the index of the maximum value in score vector. Meanwhile, considering the challenge of optimizing the MLP layer, the entire network is trained by a multi-loss collaborative optimization strategy as shown in Figure 5.

4. Experiments

4.1. Implementation Details

The proposed method is implemented with Pytorch. The input sequences are randomly/center cropped into 224×224 during training/inference. We employ SGD as the optimizer with the weight decay of 0.0003 and momentum of 0.9. The learning rate is linearly ramped up to 0.01 during the first 3 epochs, and then decayed with a cosine schedule [23]. The training lasts for 100 epochs. The data augmentation only includes random clipping and rotation. Similar to [43], all of our experiments except NTU-RGBD are pre-trained on 20BN Jester V1 dataset [24]. Moreover, three sub-branches are configured in DTN. The number of spatial and temporal feature learning blocks in DSN and DTN are $M = 6$ and $N = 6$, respectively. We refer to this setting as the basic configuration of our network, unless otherwise specified.

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

The proposed method achieves state-of-the-art (SOTA) performance on four gesture and action datasets. It is noted that we only list some of the SOTA methods for comparison, and more comparison results can be found in the supplementary material.

Method	Modality	Accuracy(%)
Transformer [11]	RGB	76.50
MTUT [1]	RGB	81.33
NAS [43]	RGB	83.61
Ours	RGB	**89.58**
Transformer [11]	Depth	83.00
MTUT [1]	Depth	84.85
NAS [43]	Depth	86.10
Ours	Depth	**90.62**
Transformer [11]	RGB+Depth	84.60
MTUT [1]	RGB+Depth	85.48
MMTM [18]	RGB+Depth	86.31
MMTM [18]	RGB-D+Flow	86.93
PointLSTM [25]	point clouds	87.90
NAS [43]	RGB+Depth	88.38
human	RGB+Depth	88.40
Ours(Multiplication)	RGB+Depth	90.89
Ours(Addition)	RGB+Depth	91.10
Ours(CAPF)	RGB+Depth	**91.70**

Table 1. Comparison of the SOTA methods on the NvGesture. Similar to [21], wherein Multiplication and Addition are the score-level feature fusion methods.

4.2.1 NvGesture Dataset

The NvGesture [26] dataset focuses on human-car interaction. It in total contains 1532 dynamic gesture videos (1050 for training and 482 for testing) in 25 classes. As can be seen in Table 1, the proposed method significantly boosts performance (RGB: $\uparrow$ 5.97%, depth: $\uparrow$ 4.52% and RGB-D: $\uparrow$ 3.3%) on this dataset for both single- and multi-modal configuration, which demonstrates its generalization ability in the field of human-computer interaction and small dataset. This might be because spatiotemporal decoupled modeling method prevents overfitting of the network to some extent.

4.2.2 THU-READ Dataset

The THU-READ [32] dataset consists 1920 videos with 40 different actions performed by 8 subjects. This dataset is challenging due to small-scale and background noise. For a fair comparison with other SOTA methods, we follow the released leave-one-split-out cross validation protocol utilized in [21]. As illustrated in Table 2, our method exceeds the SOTA results and achieves the best average accuracy (87.40%) in this protocol, which further demonstrates that

our method is also robust to complicated background. We conjecture that this is mainly attributed to the proposed FRP module.

Method	Modality	Accuracy(%)
SlowFast [13]	RGB	69.58
NAS [43]	RGB	71.25
Trear [21]	RGB	80.42
Ours	RGB	**81.25**
SlowFast [13]	Depth	68.75
NAS [43]	Depth	69.58
Trear [21]	Depth	76.04
Ours	Depth	**77.92**
SlowFast [13]	RGB+Depth	76.25
NAS [43]	RGB+Depth	78.38
Trear [21]	RGB+Depth	84.90
Ours(Multiplication)	RGB+Depth	86.10
Ours(Addition)	RGB+Depth	86.25
Ours(CAPF)	RGB+Depth	**87.04**

Table 2. Comparison of the SOTA methods on the THU-READ.

4.2.3 IsoGD Dataset

The Chalearn IsoGD [35, 36] dataset contains 47,933 RGB-D gesture videos divided into 249 kinds of gestures and is performed by 21 individuals. It is a much harder dataset because (1) it covers gestures in multiple fields and different motion scales from subtle fingertip movements to large arm swings, and (2) many gestures have a high similarity. However, as shown in Table 3, our method also performs well on this dataset, possibly due to the hierarchical and compact features learned from the multi-scale network with the recoupling structure that can capture subtle differences from similar gestures.

4.2.4 NTU-RGBD Dataset

The NTU RGB-D [28] is a large-scale human action dataset, which contains more than 56,000 multi-view videos of 60 actions performed by 40 subjects. This dataset is challenging due to large intra-class and viewpoint variations. As the skeleton information is available, many recent works tend to perform 2D/3D skeleton-based action recognition on this dataset since skeleton inherently highlights the key information of human body, whilst being robust to various illuminations and complex backgrounds. However, the generalization ability and robustness of skeleton-based methods are limited. In this paper, we only use the modalities of color and depth for action recognition. As shown in Table 4, we achieve the SOTA performance on both released protocols: Cross-view (CV) and Cross-subject (CS). Comparing

Method	Modality	Accuracy(%)
3DDSN [9]	RGB	46.08
AttentionLSTM [47]	RGB	57.42
NAS [43]	RGB	58.88
Ours	RGB	**60.87**
AttentionLSTM [47]	Depth	54.18
3DDSN [9]	Depth	54.95
NAS [43]	Depth	55.68
Ours	Depth	**60.17**
AttentionLSTM [47]	RGB+Depth	61.05
NAS [43]	RGB+Depth	65.54
Ours(Multiplication)	RGB+Depth	66.71
Ours(Addition)	RGB+Depth	66.68
Ours(CAPF)	RGB+Depth	**66.79**

Table 3. Comparison of the SOTA methods on the IsoGD.

Method	Modality	CS(%)	CV(%)
Directed-GNN [30]	Skeleton	89.9	96.1
Shift-GCN [6]	Skeleton	90.7	96.5
DC-GCN+ADG [5]	Skeleton	90.8	96.6
CTR-GCN [4]	Skeleton	92.4	96.8
Chained Multi-stream [48]	RGB	80.8	-
SLTEP [2]	Depth	58.2	-
DynamicMaps+CNN [38]	Depth	87.1	84.2
DSSCA-SSLM [29]	RGB+Depth	74.9	-
Cooperative CNN [40]	RGB+Depth	86.4	89.1
Deep Bilinear [17]	RGB-D+Skeleton	85.4	90.7
P4Transformer [12]	point	90.2	96.4
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 of the SOTA methods on the NTU-RGBD.

with the skeleton-based SOTA method CTR-GCN [4], the proposed method achieves about 2% improvement on CS protocol and 0.5% on CV protocol. The performance of depth modality can be on par or even better, which further demonstrates the robustness of our method to noisy background and its strong motion perception abilities.

5. Ablation Study

NvGesture and THU-READ(CS2) are employed for the ablation study. All of experiments are conducted on RGB data modality except Sec.5.3. We refer the reader to supplementary material for more ablation studies.

FRP	Multi-loss	Sharpness	k-NN Attention	Accuracy(%) Nv	Accuracy(%) THU
×	×	×	×	85.21	75.24
✓	×	×	×	86.67	78.33
✓	✓	×	×	87.08	80.53
✓	✓	✓	×	89.13	80.91
✓	✓	✓	✓	**89.58**	**81.67**

Table 5. Impacts of some introduced components. The Multi-loss means that we impose constraint loss on each sub-branch in DTN.

5.1. Impacts of Embedded Components

In Table 5, we show the impacts by introducing different components on the proposed network. First, we observe that in the absence of FRP module, the network does not work well either on NvGeture or THU-READ datasets, which shows that the early guidance of the network to focus on some local significant regions is beneficial to prevent the model from being trapped into the local optimum. We also demonstrate its robustness to lighting by visualization in supplementary material. In addition, imposing a constraint to each branch of the network can prevent the model from overfitting to one of the branches. Sharpening of the output distribution can encourage each sub-branch to learn more discriminative features. Finally, removing some of redundant information through k-NN Attention in the temporal representation can also bring certain performance gains.

5.2. Effect of Recoupling Representation Learning

As shown in Figure 6 (a), recoupling learning can boost the performance (Nv: $\uparrow 3\%$, THU: $\uparrow 2\%$). This is because distilling the knowledge from the temporal domain to enhance the spatial representation can help the network focus more on the informative features during training. The idea behind it may be that the inner loop optimization mechanism based on self-distillation can help the network deviate from the local optima as soon as possible and move toward the global optimal solution. Figure 6 (b) shows the influence of distillation temperature $\mathcal{T}$ on network performance. We observe that different data domains enjoy different temperatures $\mathcal{T}$ during training: 0.4 for gesture dataset and 0.5 for action dataset.

Dataset	Baseline (Add)	CmSI	CmTI	CmSTI
NvGesture	91.10	91.32	91.53	91.70
THU-READ	86.35	87.34	88.75	90.00

Table 6. Effects of cross-modal spatiotemporal information interaction. CmSI: Cross-modal Spatial information Interaction only. CmTI: Cross-modal Temporal information Interaction only. CmSTI: Cross-modal spatiotemporal information Interaction.

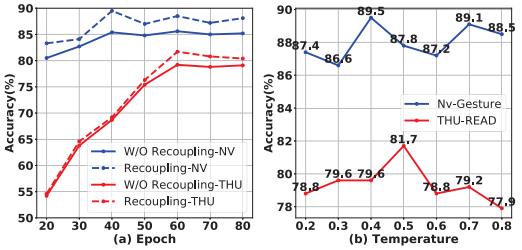

Figure 6. The ablation study of the recoupling learning. (a) The effect of recoupling strategy on network performance. (b) The effect of distillation temperature on network performance.

5.3. Cross-modal Information Interaction

As shown in Table 6, the communication of spatial or temporal information can boost the performance, wherein the information exchange at the spatial or temporal level alone can also bring performance gains, which reflects that the interaction of cross-domain knowledge can help learn discriminative features. The last experiment shows that the joint effective cross-domain information interaction at the spatial and temporal levels can bring the largest performance gains (Nv: $\uparrow 0.6\%$, THU: $\uparrow 3.7\%$), verifying that multi-modal feature learning can benefit from spatiotemporal independent information transformation.

6. Conclusion

We propose a method for unimodal decoupling and recoupling learning as well as a cross-modal interactive learning and fusion. Firstly, we observe that spatiotemporal recoupling learning is very effective as it can lower the optimization difficulty, especially under small dataset settings. Secondly, guiding the network to focus on the local important areas helps boost the performance. Finally, we prove that interacting with the cross-modal spatiotemporal information in two independent dimensions of space and time, respectively, can encourage the network to extract and fuse multi-modal spatiotemporal features.

7. Acknowledgment

This work was supported by the Alibaba Group through Alibaba Research Intern Program, the External cooperation key project of Chinese Academy Sciences 173211KYSB20200002, the Chinese National Natural Science Foundation Projects 61876179 and 61961160704, the Science and Technology Development Fund of Macau (0025/2019/AKP, 0008/2019/A1, 0010/2019/AFJ, 0004/2020/A1 0070/2020/AMJ), Guangdong Provincial Key R&D Programme: 2019B010148001, the InnoHK program, and the open Research Projects of Zhejiang Lab No. 2021PF0AB01.

References

[1] Mahdi Abavisani, Hamid Reza Vaezi Joze, and Vishal M. Patel. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 6

[2] Ji Xiaopeng A B C, Cheng Jun A B C, Tao Dapeng D, Wu Xinyu A B C, and Feng Wei A B C. The spatial laplacian and temporal energy pyramid representation for human action recognition using depth sequences - sciencedirect. *Knowledge-Based Systems*, 122(C):64–74, 2017. 7

[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 (CVPR)*, July 2017. 1, 3

[4] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. 7

[5] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *Computer Vision–ECCV 2020: 16th European Conference*, pages 536–553. Springer, 2020. 7

[6] K. Cheng, Y. Zhang, X. He, W. Chen, and H. Lu. Skeleton-based action recognition with shift graph convolutional network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[7] Alban Main De Boissiere and Rita Noumeir. Infrared and 3d skeleton feature fusion for rgb-d action recognition. *IEEE Access*, 8:168297–168308, 2020. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[9] Jiali Duan, Jun Wan, Shuai Zhou, Xiaoyuan Guo, and Stan Li. A unified framework for multi-modal isolated gesture recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14:1–16, 02 2018. 7

[10] Andrea D'Eusanio, Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. A transformer-based network for dynamic hand gesture recognition. In *2020 International Conference on 3D Vision (3DV)*, pages 623–632, 2020. 1

[11] Andrea D'Eusanio, Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. A transformer-based network for dynamic hand gesture recognition. In *2020 International Conference on 3D Vision (3DV)*, pages 623–632. IEEE, 2020. 6

[12] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14204–14213, June 2021. 7

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 7

[14] Dongliang He, Zhichao Zhou, Chuang Gan, Fu Li, Xiao Liu, Yandong Li, Limin Wang, and Shilei Wen. Stnet: Local and global spatial-temporal modeling for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8401–8408, 2019. 3

[15] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[17] Jian-Fang Hu, Wei-Shi Zheng, Jiahui Pan, Jianhuang Lai, and Jianguo Zhang. Deep bilinear learning for rgb-d action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 7

[18] 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. 6

[19] M Esat Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *European Conference on Computer Vision*, pages 731–747. Springer, 2020. 1

[20] Yu Kong and Yun Fu. Bilinear heterogeneous information machine for rgb-d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1054–1062, 2015. 3

[21] 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*, 2021. 1, 6, 7

[22] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Decoupled spatial-temporal transformer for video inpainting. *arXiv preprint arXiv:2104.06637*, 2021. 3

[23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[24] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 6

[25] Yuecong Min, Yanxiao Zhang, Xiujuan Chai, and Xilin Chen. An efficient pointlstm for point clouds based gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5761–5770, 2020. 6

[26] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4207–4215, 2016. 2, 6

[27] Xuan Son Nguyen, Luc Brun, Olivier Lézoray, and Sébastien Bougleux. A neural network based on spd manifold learning for skeleton-based hand gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12036–12045, 2019. 1

[28] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 3, 7

[29] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1045–1058, 2018. 7

[30] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019. 7

[31] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3

[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. 3, 6

[33] 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. 1

[34] Jun Wan, Guodong Guo, and Stan Z 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. 1

[35] 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*, pages 1–12, 2020. 7

[36] 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 the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–64, 2016. 3, 7

[37] Huogen Wang, Pichao Wang, Zhanjie Song, and Wanqing Li. Large-scale multimodal gesture recognition using heterogeneous networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[38] P. Wang, W. Li, Z. Gao, C. Tang, and P. O. Ogunbona. Depth pooling based large-scale 3-d action recognition with convolutional neural networks. *IEEE Transactions on Multimedia*, 2018. 1, 7

[39] Pichao Wang, Wanqing Li, Zhimin Gao, Yuyao Zhang, Chang Tang, and Philip Ogunbona. Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2017. 3

[40] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu. Cooperative training of deep aggregation networks for rgb-d action recognition. 2017. 7

[41] 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. 1, 3

[42] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Wen Xie, Hao Li, and Rong Jin. Kvt: k-nn attention for boosting vision transformers. *arXiv preprint arXiv:2106.00515*, 2021. 4

[43] 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, 3, 6, 7

[44] Zehua Zhang and David Crandall. Hierarchically decoupled spatial-temporal contrast for self-supervised video representation learning. *arXiv preprint arXiv:2011.11261*, 2020. 3

[45] Benjia Zhou, Yunan Li, and Jun Wan. Regional attention with architecture-rebuilt 3d network for rgb-d gesture recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3563–3571, May 2021. 1, 2, 4

[46] Guangming Zhu, Liang Zhang, Lin Mei, Jie Shao, Juan Song, and Peiyi Shen. Large-scale isolated gesture recognition using pyramidal 3d convolutional networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 19–24. IEEE, 2016. 3

[47] 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. 2, 7

[48] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2904–2913, 2017. 7