Searching Multi-Rate and Multi-Modal Temporal Enhanced Networks for Gesture Recognition

Zitong Yu, Benjia Zhou, Jun Wan, *Senior Member, IEEE*, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z. Li, *Fellow, IEEE* and Guoying Zhao, *Senior Member, IEEE*

***Abstract*—Gesture recognition has attracted considerable attention owing to its great potential in applications. Although the great progress has been made recently in multi-modal learning methods, existing methods still lack effective integration to fully explore synergies among spatio-temporal modalities effectively for gesture recognition. The problems are partially due to the fact that the existing manually designed network architectures have low efficiency in the joint learning of multi-modalities. In this paper, we propose the first neural architecture search (NAS)-based method for RGB-D gesture recognition. The proposed method includes two key components: 1) enhanced temporal representation via the proposed 3D Central Difference Convolution (3D-CDC) family, which is able to capture rich temporal context via aggregating temporal difference information; and 2) optimized backbones for multi-sampling-rate branches and lateral connections among varied modalities. The resultant multi-modal multi-rate network provides a new perspective to understand the relationship between RGB and depth modalities and their temporal dynamics. Comprehensive experiments are performed on three benchmark datasets (IsoGD, NvGesture, and EgoGesture), demonstrating the state-of-the-art performance in both single- and multi-modality settings. The code is available at https://github.com/ZitongYu/3DCDC-NAS.**

***Index Terms*—3D-CDC, NAS, RGB-D gesture recognition.**

I. Introduction

As one of the multi-modal video understanding problems, RGB-D based video gesture recognition [1], [2], [3], [4] has been applied to many real-world applications, such as virtual reality [5] and human-computer interaction [6]. 3D convolutional neural network (3DCNN) [7], [8], [9], [10] and long short-term memory (LSTM) [11] have been adopted to learn the spatial-temporal features for gesture recognition. However, the learned spatio-temporal representation is still easily contaminated by irrelevant factors (e.g., illumination and background). A feasible solution is to add an extra enhanced temporal feature learning module [12], [13], [14], which is computational costly and tricky for the off-the-shelf 3DCNNs. **How to learn efficient spatio-temporal features in basic convolution operator for enriching temporal context is worth exploring for gesture recognition.**

As gestures have various temporal ranges, modeling such visual tempos would benefit for gesture recognition. Previous methods [15], [16], [17] attempt to construct the frame pyramid for such purpose, with each branch of the frame pyramid sampling the input frames at a different rate. However, the architecture (i.e., network structure) of each branch and relations (i.e., lateral connections) among the multi-rate branches are usually shared and hand-designed, which is sub-optimal for message propagation. Hence, **how to discover better-suited architectures and lateral connections among multi-rate branches is crucial.**

For RGB-D based gesture recognition, complementary feature learning from different data modalities is beneficial. For example, the depth data is easy to distinguish foregrounds (i.e., face, hands, and arms) from backgrounds while RGB data provides detailed texture/color appearances. However, most existing methods [7], [8], [18], [19], [20] conduct the multi-modal fusion via coarse strategies (e.g, score fusion or last layer concatenation), which may leverage the multi-modal information insufficiently. Thus, **to design more reasonable multi-modal fusion strategy is not a trivial work.**

Motivated by the above observations, we propose a novel spatio-temporal convolution family called 3D Central Difference Convolution (3D-CDC), to exploit the rich local motion and enhance the spatio-temporal representation. Moreover, over the 3D-CDC-based enhanced search space, Neural Architecture Search (NAS) is adopted to automatically discover the optimized multi-rate and multi-modal networks for RGB-D gesture recognition. Our contributions include:

- A novel spatio-temporal convolution family, 3D-CDC, is proposed, intending to capture rich temporal context via aggregating temporal difference information. Without introducing extra parameters, 3D-CDC can replace the vanilla 3D convolution, and plug and play in existing 3DCNNs for various modalities with enhanced temporal modeling capacity.
- We propose a two-stage NAS method to automatically discover well-suited backbones and lateral connections for the multi-rate and multi-modal networks, which effectively explores RGB-D-temporal synergies and represents global dynamics.
- To our best knowledge, this is the first approach that searches multi-rate and multi-modal architectures for RGB-D gesture recognition. Our searched architecture

The first two authors contribute equally. Corresponding authors: Jun Wan {jun.wan}@nlpr.ia.ac.cn, and Guoying Zhao {guoying.zhao}@oulu.fi.

Z. Yu, H. Chen, X. Liu and G. Zhao are with Center for Machine Vision and Signal Analysis, University of Oulu, Oulu 90014, Finland.

B. Zhou is with Macau University of Science and Technology, Macau 999078, China.

J. Wan is with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

P. Wang is with DAMO Academy, Alibaba Group (U.S.) Inc., Bellevue, WA, 98004, USA.

Stan Z. Li is with Westlake University, Hangzhou 310012, China.

Manuscript received August 21, 2020.

provides a new perspective to understand the relationship among multi-rate branches as well as modalities.
- Our proposed method achieves state-of-the-art (SOTA) performance on three benchmark datasets on both single- and multi-modality testing.

In the rest of the paper, Sec. II provides the related work. Sec. III formulates the 3D-CDC family, and introduce the two-stage multi-rate and multi-modal NAS algorithm. Sec. IV provides rigorous ablation studies and evaluates the performance of the proposed models on three benchmark datasets. Sec. V shows the visualization results and discusses transferability to the action recognition task. Finally, a conclusion with future directions is given in Sec. VI.

II. Related Work

In this section, we first introduce some recent progress in multi-modal gesture recognition. Then, previous video-based NAS methods will be reviewed.

Multi-Modal Gesture Recognition. For video-based gesture recognition, it is challenging to track the motion of hands and arms owing to the large degree of freedom. Many hand-crafted feature based [21], [22], [23], [24] and deep learning-based methods [25], [26], [11], [27], [28] are proposed to tackle this issue. As for the learning-based gesture recognition, on one hand, 3DCNNs including C3D [29], Res3D [30], I3D [31] and SlowFast [15] are utilized for gesture feature extractor. On the other hand, LSTM variants such as Atten-ConvLSTM [32], [32] and PreRNN [33] are introduced for temporal memory propagation. Based on the CNNs, several extended modules [12], [13], [14], [34] and convolution operators [35], [36] are developed for enhancing the spatio-temperal representation. However, most of them need extra structures and learnable parameters to modulate the original spatio-temporal features. In this paper, we propose 3D-CDC for modeling a rich temporal context, which is vital for describing fine-grained hands/arms motion. Among these methods, Lee et al.'s [13] motion feature network (MFNet) and Yu et al.'s [36], [37] central difference convolution (CDC) are the most similar to our work. Instead of the fixed motion filter used in MFNet and only the spatial context in CDC, our 3D-CDC learns the temporal gradient (motion) filters automatically in a unified 3D convolution operator.

Due to the development of the RGB-D cameras like Intel RealSense, RGB and depth modalities [8], [7], [38] are favorite to complementarily fuse and improve the performance. Besides the RGB-D data, some other modalities such as optical flow [39], [7], depth flow [38], skeleton [40], [41] and saliency video [42], [43], [9], [43] are also employed for gesture recognition. However, the flow and saliency modalities need extra computational cost and might lose some vital information after pre-processing.

In terms of the multi-modal fusion strategy, decision-level fusion [44], [9], [45] and feature-level fusion [7], [8], [38], [20] methods have developed for integrating mutual knowledge from varied modalities. Despite achieving SOTA performance, the existing multi-modal fusion strategies for gesture recognition are designed manually and coarsely, which might be sub-optimal for message propagation between modalities. In this paper, we prefer to discover well-suitable multi-modal fusion strategies automatically via NAS.

Video-based Neural Architecture Search. Our work is motivated by recent researches on NAS [46], [47], [48], [49], [50], [51], [52], while we focus on searching for multi-rate and multi-modal networks specially for RGB-D gesture recognition. The existing NAS methods could be summarized in three categories: 1) Reinforcement learning-based [49], [50], 2) Evolution-based [53], [54], and 3) Gradient-based [47], [51], [55]. From the perspective of NAS based video classification applications, single-modal based [56], [57], [58] and multi-modal based [59], [60] methods have been developed for the action recognition task. Unlike AssembleNet [60] which searches on RGB and optical flow modalities with single frame rate inputs, our work focuses on searching the synergies among multi-rate and multi-modal branches.

In terms of the search space design, cell-based NASNet [50] space is favorite due to its flexibility and rich capacity. Besides, some novel operators, such as extended convolution [36], [61] and attention [58], [62], are introduced into search space, which is proved to be beneficial for searching more powerful architecture. However, there are still no operators specially designed for temporal enhancement.

To our best knowledge, no NAS based method has ever been proposed for RGB-D gesture recognition. To fill in the blank, we search multi-rate and multi-modal networks over the temporal enhanced search space for RGB-D gesture recognition.

III. Methodology

In this section, we first introduce the 3D-CDC family in Sec. III-A. Over the 3D-CDC based space, we then propose the two-stage multi-rate and multi-modal NAS in Sec. III-B.

A. Temporal Enhancement via 3D-CDC

In classical 3DCNNs, 3D convolution is the most fundamental operator for spatio-temporal feature representation. In this subsection, for simplicity, the 3D convolutions are described in 3D (without channel) while an extension to 4D is straightforward. There are two main steps in the vanilla 3D convolution: 1) *sampling* local receptive field cube $\mathcal{C}$ over the input feature map x; 2) *aggregation* of sampled values via weighted summation with learnable weights w. Hence, the output feature map $Vanilla$ can be formulated as

$$Vanilla(p_0) = \sum_{p_n \in \mathcal{C}} w(p_n) \cdot x(p_0 + p_n), \tag{1}$$

where p_0 denotes current location on both input and output feature maps while p_n enumerates the locations in $\mathcal{C}$. In this subsection, 3D convolution with kernel size $3\times3\times3$ and dilation 1 is used for demonstration, and the other configurations are analogous. The local receptive field cube for the 3D convolution is $\mathcal{C}$=$\{(-1,-1,-1),(-1,-1,0),\cdots,(0,1,1),(1,1,1)\}$.

Spatio-Temporal Central Difference Convolution (3D-CDC-ST). Inspired by the CDC [36] which introduces spatial gradient cues into representation learning, we integrate spatio-temporal gradient information into a unified 3D convolution operator. It is worth noting that such spatial gradient

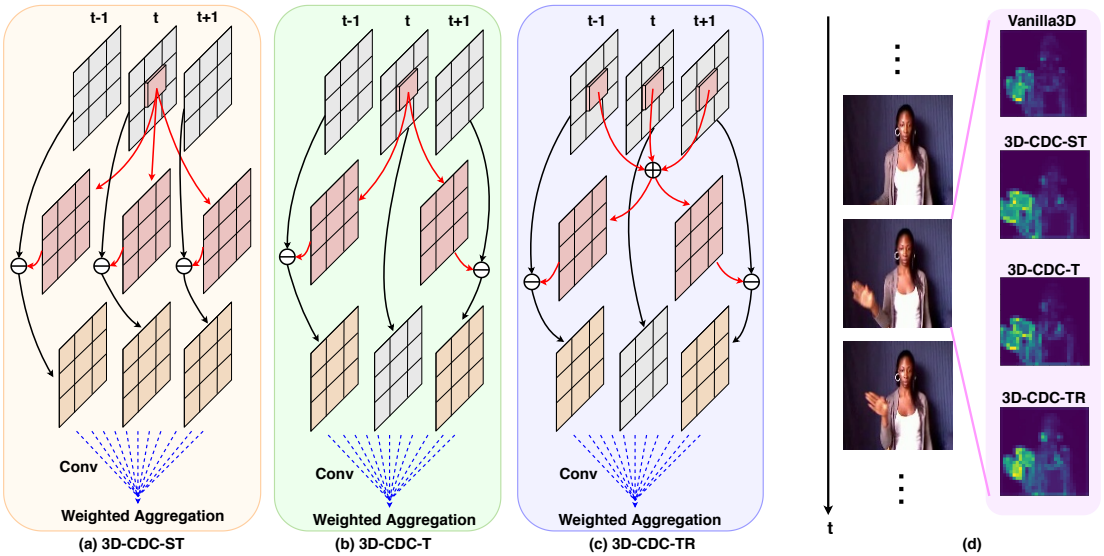


Fig. 1: Our proposed 3D-CDC family with kernel size 3, which could be used as novel operators for NAS. (a) 3D-CDC-ST considers the central difference information in the whole local spatio-temporal regions. (b) 3D-CDC-T only calculates the central difference clues from the local spatio-temporal regions of the adjacent frames. (c) 3D-CDC-TR is similar to 3D-CDC-T but adopts the temporal central mean pooling before calculating the central difference term, which is more robust to temporal noise. The symbols $\ominus$ and $\oplus$ denote element-wise subtraction and mean operations, respectively. (d) Feature response of various convolutions in RGB modality. Compared with vanilla 3D convolution, the 3D-CDC family enhances the temporal context obviously.

and temporal difference designs are widely used in the dense optical flow [63] calculation. Instead of the optical flow only performed in RGB sequence level, networks with stacked spatio-temporal CDC would be regularized to learn more local motion context in both RGB sequence and deep feature level, which is able to model fine-grained temporal dynamics for gesture recognition.

Similarly, spatio-temporal CDC also consists of two steps, i.e., *sampling* and *aggregation*. The sampling step is similar to vanilla 3D convolution but the aggregation step is different: as illustrated in Fig. 1(a), spatio-temporal CDC prefers to aggregate the center-oriented spatio-temporal gradient of sampled values. Eq. (1) becomes

$$CDC(p_0) = \sum_{p_n \in \mathcal{C}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)). \tag{2}$$

When p_n=(0,0,0), the gradient value always equals to zero with respect to the central location p_0 itself. For the gesture recognition task, both spatio-temporal intensity-level semantic information and gradient-level difference message are crucial and complementary. The former one is good at global modeling and robust to sensor-based noise while the latter one focuses more on local appearance and motion details and might be influenced by noise. As a result, combining vanilla 3D convolution with 3D-CDC might be a feasible manner to provide more robust and discriminative modeling capacity. Therefore we generalize spatio-temporal CDC as

$$\begin{aligned} CDC_{ST}(p_0) &= \theta \cdot CDC(p_0) + (1-\theta) \cdot Vanilla(p_0) \\ &= \underbrace{\sum_{p_n \in \mathcal{C}} w(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla 3D convolution}} + \theta \cdot \underbrace{(-x(p_0) \cdot \sum_{p_n \in \mathcal{C}} w(p_n))}_{\text{spatio-temporal CD term}}, \end{aligned} \tag{3}$$

where hyperparameter $\theta \in [0, 1]$ tradeoffs the contribution between intensity-level and gradient-level information. Please note that $w(p_n)$ is shared between vanilla 3D convolution and spatio-temporal central difference (CD) term, thus no extra parameters are added.

Temporal Central Difference Convolution (3D-CDC-T). Unlike the aforementioned spatio-temporal CDC considering both spatial and temporal gradient cues, we propose a version with only temporal central differences. As shown in Fig. 1(b), the sampled local receptive field cube $\mathcal{C}$ is separated into two kinds of regions: 1) the region in the current time step $\mathcal{R}'$, and 2) the regions in the adjacent time steps $\mathcal{R}''$. In the setting of a temporal CDC, the central difference term is only calculated from $\mathcal{R}''$. Thus the generalized temporal CDC can be formulated via modifying Eq. (3) as

$$CDC_T(p_0) = \underbrace{\sum_{p_n \in \mathcal{C}} w(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla 3D convolution}} + \theta \cdot \underbrace{(-x(p_0) \cdot \sum_{p_n \in \mathcal{R}''} w(p_n))}_{\text{temporal CD term}}. \tag{4}$$

Temporal Robust Central Difference Convolution (3D-CDC-TR). In consideration of the sensor noise especially in the depth modality, we also propose a version with the temporal robust central difference. Similarly, the temporal robust CDC only calculates the difference term from the regions in the adjacent time steps $\mathcal{R}''$. As shown in Fig. 1(c), the robust temporal center is represented by averaging the spatial centers of all time steps (i.e., p_0^{t-1}, p_0 and p_0^{t+1}) within $\mathcal{C}$. The robust temporal center-oriented gradient might be less sensitive to the pixel jitters from the adjacent time steps. The generalized temporal robust CDC can also be formulated via

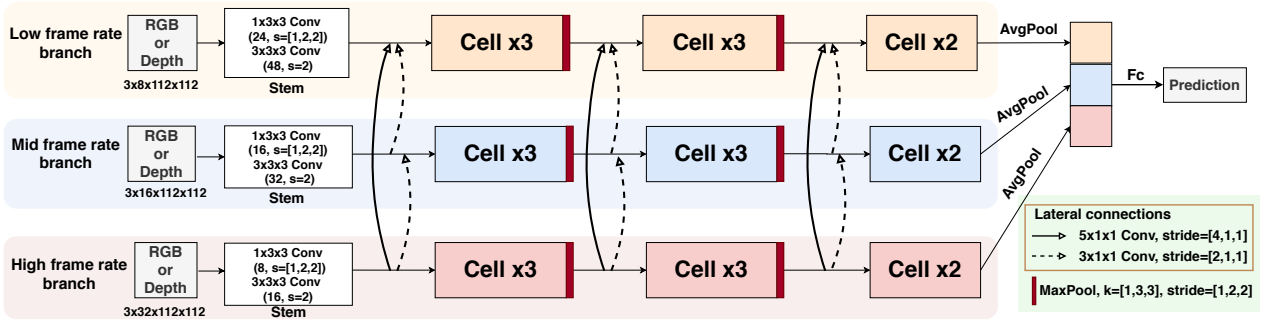


Fig. 2: Architecture search space in the first stage. Single-modal (RGB or depth) multi-rate frames are adopted as inputs. Here we utilize 3 branches with different frame rates (e.g., uniformly sampling the original videos into 8, 16, and 32 frames, respectively), and it also can be extended to more branches according to the actual situation. In this search stage, inspired by [15], the architectures of all lateral connections are fixed with temporal convolutions. And the outputs of the lateral connections are concatenated with the features from the target branch. The channel numbers are doubled after each MaxPool layer. The architecture of the cells from multi-rate branches to be searched can be shared or unshared (see Sec. IV-C for ablation study).

modifying Eq. (3) as

$$CDC_{TR}(p_0) = \underbrace{\sum_{p_n \in \mathcal{C}} w(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla 3D convolution}} + \underbrace{\theta \cdot (-Avg[x(p_0^{t-1}), x(p_0), x(p_0^{t+1})] \cdot \sum_{p_n \in \mathcal{R}''} w(p_n)) .}_{\text{temporal robust CD term}} \tag{5}$$

We will henceforth refer to these three generalized versions (i.e., Eq. (3), (4) and (5)) as 3D-CDC-ST, 3D-CDC-T and 3D-CDC-TR, respectively. The ablation studies about the 3D-CDC family and hyperparameter θ are conducted in Sec. IV-C.

Relation between 3D-CDC and 3D Vanilla Convolution. Compared to 3D vanilla convolution, 1) 3D-CDC-ST regularizes the spatio-temporal representation with more detailed spatial cues and temporal dynamics, which might be suitable for scene-aware video understanding tasks; 2) 3D-CDC-T enhances the spatio-temporal representation with only rich temporal context, which might be assembled for video temporal reasoning tasks; and 3) 3D-CDC-TR introduces robust but slighter temporal evolution cues for spatio-temporal representation, which might perform robustly even in noisy scenarios. In particular, 3D-CDC-ST, 3D-CDC-T, and 3D-CDC-TR degrade to vanilla 3D convolution when θ=0. As illustrated in Fig. 1(d), the 3D-CDC family provides more details about the trajectory of the left arm, and such a local motion context is crucial to gesture recognition. More visualizations are shown in Sec. V-C.

B. Multi-Rate and Multi-Modal NAS

In order to seek the best-suited backbones and lateral connections for multi-rate and multi-modal networks, we propose a two-stage NAS method to 1) search backbones for multi-rate single-modal networks first, and then 2) search lateral connections for multi-modal networks based on the searched backbones. The iterative procedure is outlined in Algorithm 1. More technical details can be referred to two gradient-based NAS methods [47], [51].

Algorithm 1 Two-Stage Multi-Rate and Multi-Modal NAS

Stage1: *Fix lateral connections, and search backbones*
For multi-rate backbones, create a mixed operation $\tilde{o}_b^{(i,j)}$ parametrized by $\alpha_b^{(i,j)}$ for each edge (i, j)
for each of modalities $\mathcal{M}$ **do**
  Fix the lateral connections among multi-rate branches
  while not converged **do**
    Update architecture α_b by descending $\nabla_{\alpha_b}\mathcal{L}_{val}(\Phi_b, \alpha_b)$
    Update weights Φ_b by descending $\nabla_{\Phi_b}\mathcal{L}_{train}(\Phi_b, \alpha_b)$
  end
  Derive the final backbone of the current modality based on the learned α_b
end
Stage2: *Fix backbones, and search lateral connections*
For lateral connections, create a mixed operation $\tilde{o}_c^{(i,j)}$ parametrized by $\alpha_c^{(i,j)}$ for each edge (i, j)
Initialize and fix the multi-rate and multi-modal backbones searched in ***Stage 1***
while not converged **do**
  Update architecture α_c by descending $\nabla_{\alpha_c}\mathcal{L}_{val}(\Phi_c, \alpha_c)$
  Update weights Φ_c by descending $\nabla_{\Phi_c}\mathcal{L}_{train}(\Phi_c, \alpha_c)$
end
Derive the final lateral connections based on the learned α_c

1) Stage 1: Searching Backbones for Multi-Rate Single-Modal Networks: In SlowFast Networks [15], low- and high-rate branches are utilized for complementarily capturing dynamic visual tempos. However, the coarse hand-defined architecture with vanilla convolutions limits its representation capacity. Here we search optimal rate-aware backbones over the temporal enhanced search space.

As illustrated in Fig. 2, in the first stage, our goal is to search for cells to form multi-rate backbones for single-modal gesture recognition. As for the cell-level structure, similar to [47], each cell is represented as a directed acyclic graph (DAG) of K nodes $\{n\}_{i=0}^{K-1}$, where each node represents a network layer. We denote the operation space as $\mathcal{O}_b$, which consists of seven designed candidate operations: '*Zero*', '*Identity*', '*Conv_1x3x3*', '*CDC-T-06_3x1x1*', '*CDC-T-06_3x3x3*', '*CDC-TR-03_3x1x1*' and '*CDC-TR-03_3x3x3*'. To be specific, '*CDC-T-06*' and '*CDC-TR-03*' denote the 3D-CDC-T with θ=0.6 and 3D-CDC-TR with θ=0.3, respectively.

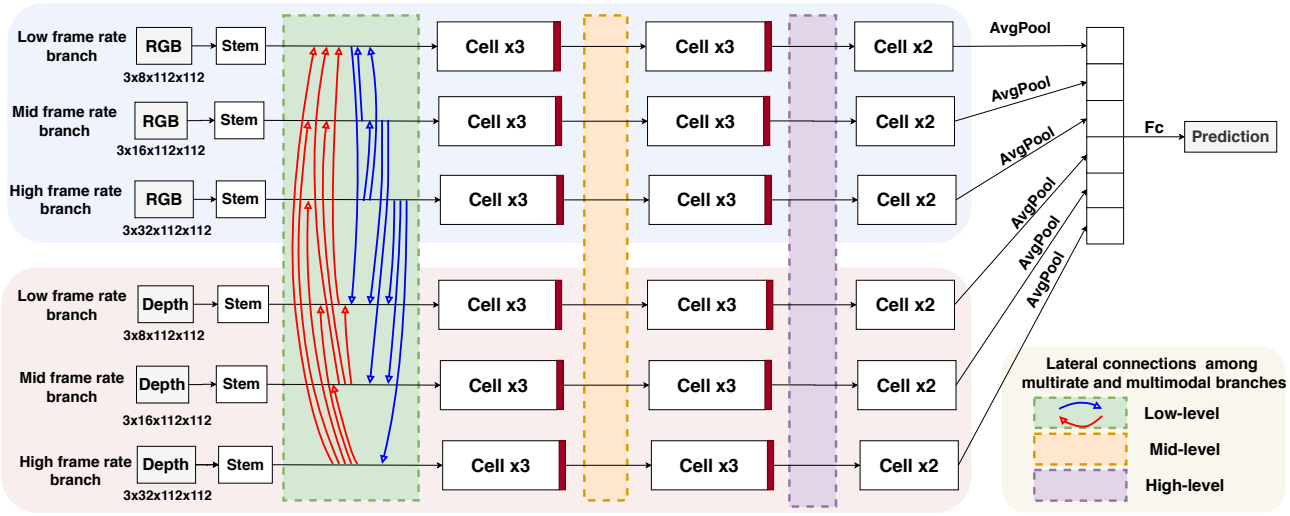


Fig. 3: Architecture search space in the second stage. Multi-modal and multi-rate frames are adopted as inputs. Here we utilize 3 branches with different frame rates for two modalities (RGB and depth), respectively. In this search stage, all cells are initialized with the searched structures in the first stage and then fixed. The architecture of the low-, mid- and high-level lateral connections to be searched can be shared or unshared.

These settings of θ are based on the ablation study results in Section IV-C. We also consider a vanilla operation space with vanilla 3D convolutions instead of 3D-CDC for comparison.

The multi-rate backbones for each of modalities $\mathcal{M}$ to be searched have the architecture parameters $\alpha_b^{(i,j)}$. Each edge (i,j) of DAG represents the information flow from node n_i to node n_j, which consists of the candidate operations weighted by the architecture parameter $\alpha_b^{(i,j)}$. Specially, each edge (i,j) can be formulated by a function $\tilde{o}_b^{(i,j)}$ where $\tilde{o}_b^{(i,j)}(n_i) = \sum_{o_b \in \mathcal{O}_b} \eta_{o_b}^{(i,j)} \cdot o_b(n_i)$. Softmax $\eta_{o_b}^{(i,j)} = \frac{exp(\alpha_{o_b}^{(i,j)})}{\sum_{o'_b \in \mathcal{O}_\lfloor} exp(\alpha_{o'_b}^{(i,j)})}$ is utilized to relax architecture parameter $\alpha_b^{(i,j)}$ into operation weight $o_b \in \mathcal{O}_b$. The intermediate node can be denoted as $n_j = \sum_{i<j} \tilde{o}_b^{(i,j)}(n_i)$. The output node n_{K-1} is depth-wise concatenation of all the intermediate nodes excluding the input nodes.

In the searching stage, cross-entropy loss is utilized for the training loss $\mathcal{L}_{train}$ and validation loss $\mathcal{L}_{val}$. Network parameters Φ_b and architecture parameters α_b are learned via solving the bi-level optimization problem:

$$\begin{aligned} &\min_{\alpha_b} \quad \mathcal{L}_{val}(\Phi_b^*(\alpha_b), \alpha_b), \\ &s.t. \quad \Phi_b^*(\alpha) = arg\min_{\Phi_b} \mathcal{L}_{train}(\Phi_b, \alpha_b). \end{aligned} \tag{6}$$

When the searching is converged, we derive the final architectures via: 1) setting $o_b^{(i,j)} = arg\,max_{o_b \in \mathcal{O}_b, o_b \neq zero} \eta_{o_b}^{(i,j)}$, and 2) for each intermediate node, choosing two incoming edges with the two largest values of $max_{o_b \in \mathcal{O}_b, o \neq zero} \eta_{o_b}^{(i,j)}$.

2) Stage 2: Searching Lateral Connections for Multi-Rate Multi-Modal Networks: The lateral connections from most existing multi-rate [15] or multi-modal [9], [7], [20] spatio-temporal networks are designed manually, which might be sub-optimal for information exchange. Here we propose to search best-suited lateral connections among rate-aware and modality-aware branches, intending to effectively explore RGB-D-temporal synergies. In the second stage, our goal is to search for lateral connections among the multi-rate and multi-modal branches. As most of the definitions and the search procedure are similar to those in the first stage, here we only show the two main differences from the first stage.

On one hand, the composition of architecture search space is different. As shown in Fig. 3 (see low-level connections for example), the lateral connections search space can be represented as a bidirectional graph of K'=6 nodes (branches) within the modalities $\mathcal{M}$. Specially, the lateral connections from the lower frame rate branches to the higher ones are not considered because we assume that the lower frame rate branches always have less information than the higher ones. Thus, there are total 18 edges (lateral connections) inside the bidirectional graph and each edge consists of the candidate operations weighted by its corresponding architecture parameter α_c. The final output of each node is the depth-wise concatenation of all outputs of the incoming edges.

On the other hand, the design of the operation space is different. The operation space for lateral connections is denoted as $\mathcal{O}_c$, which consists of two parts: 1) '*Zero*', '*Conv_5x1x1*', '*CDC-T-06_5x1x1*', '*CDC-TR-03_5x1x1*' with $stride$=(4,1,1) for the edges from the high frame rate branches to the low frame rate branches; and 2) '*Zero*', '*Conv_3x1x1*', '*CDC-T-06_3x1x1*', '*CDC-TR-03_3x1x1*' with $stride$=(2,1,1) for the others. Specially, $stride$=1 is utilized for edges between the branches of different modalities with same frame rate. When the search is converged, for each edge (i,j), only the operation in $\mathcal{O}_c$ with the largest $\alpha_c^{(i,j)}$ is adopted. With the two-stage multi-rate and multi-modal NAS in Algorithm 1, both the final backbones and lateral connections are derived.

IV. Experiments

In this part, we first give details for benchmark datasets and experimental setup. Then, we thoroughly evaluate the impacts of 3D-CDC family, multi-rate configuration, and two-stage

NAS. Finally, we evaluate and compare our results with state-of-the-art methods on three benchmark datasets.

A. Datasets

We evaluate our method on three widely used RGB-D gesture datasets: Chalearn IsoGD [2], [64], NVGesture [1] and EgoGesture [4] datasets. The **Chalearn IsoGD dataset** [2] contains 47,933 RGB-D gesture videos divided into 249 kinds of gestures and is performed by 21 individuals. The dataset contains 35878 training, 5784 validation, and 6271 testing samples. We follow the SOTA methods [45], [32], [38] to evaluate performance on the validation set. The **NVGesture dataset** [1] focuses on touchless driver controlling. It contains 1532 dynamic gestures fallen into 25 classes. It includes 1050 samples for training and 482 for testing. The videos are recorded with three modalities (RGB, depth, and infrared). For fair evaluations with SOTA methods, infrared modality is not used in our experiments. The **EgoGesture dataset** [4] is a large multi-modal egocentric hand gesture dataset. It contains 24,161 hand gesture clips of 83 classes of gestures, performed by 50 subjects. Videos in this dataset are captured with an Intel RealSense SR300 device in RGB-D modalities across multiple indoor and outdoor scenes.

B. Implementation Details

Our proposed method is implemented with Pytorch. Cell nodes K=4 and K'=6 are used as the default setting. The optical flow is extracted by pyflow [65] - a python wrapper for dense optical flow [63]. **In the search phase**, partial channel connection and edge normalization [51] are adopted. The initial channel numbers for low, mid, and high frame rate branches are 24, 16, and 8, respectively, which double after searching. There are 8 cells for each branch in the search stage, which increases to 12 cells after searching. SGD optimizer with learning rate lr=1e-2 and weight decay wd=5e-5 is utilized when training the network weights. The architecture parameters are trained with Adam optimizer with lr=6e-4 and wd=1e-3. The lr decays with factor 0.5 in the 20^{th} epoch. We search 30 epochs on the training set of IsoGD dataset [2] with batch size 20 while architecture parameters are not updated in the first 10 epochs. Especially, $\mathcal{L}_{train}$ is calculated on the first half of the training set while $\mathcal{L}_{val}$ on the latter part. The whole two-stage NAS costs 12 days on four P100s. **In the training phase**, models are trained with SGD optimizer with initial lr=1e-2 and wd=5e-5. The lr decays with factor 0.1 when the validation accuracy has not improved within 3 epochs. Random horizontal flip and spatial crops are utilized for data augmentation. We train models with batch size 12 for maximum 80 epochs on four RTX-2080Ti GPUs.

C. Ablation Study

All ablation studies are trained from scratch and evaluated on the validation set of the IsoGD dataset.

Impact of 3D-CDC for Modalities. In these experiments, we use C3D [29] as the backbone and sequence size 3×16×112×112 as the inputs. According to Eq. (3), (4) and (5), θ controls the contribution of the temporal difference cues. As illustrated in Fig. 4, 3D-CDC-T improves the accuracy

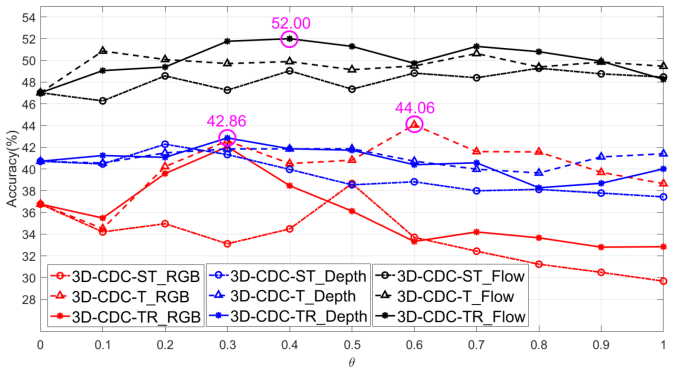


Fig. 4: Impact of 3D-CDC for RGB, depth and optical flow.

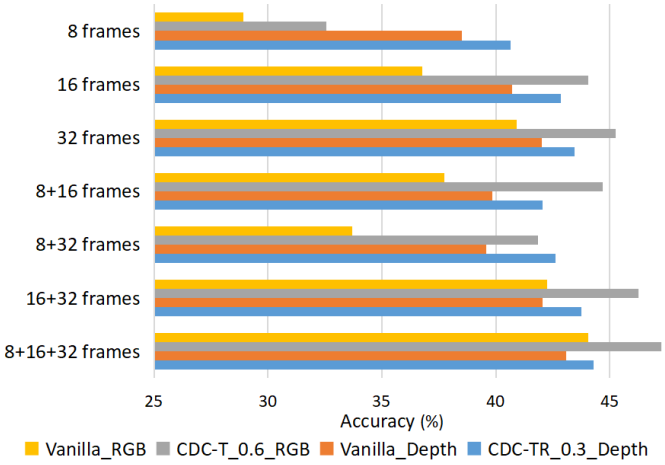


Fig. 5: The ablation study of the multi-rate networks. '8+16+32 frames' means that there are three branches with different frame rate (i.e., temporally downsampling to 8, 16 and 32 frames, respectively) as inputs. 'CDC-T_0.6_RGB' and 'CDC-TR_0.3_Depth' denotes using 3D-CDC-T with θ=0.6 for RGB inputs and 3D-CDC-TR with θ=0.3 for depth inputs, respectively.

of RGB modality dramatically. Compared with vanilla 3D convolution (i.e., θ=0), 3D-CDC-T gains 7.3% when θ=0.6, which indicates the advantages of temporal difference context. One highlight is that, assembled with 3D-CDC-T, the RGB modality is able to obtain comparable accuracy (44.06% vs. 47.02%) with optical flow modality, indicating an excellent dynamic modeling capacity of 3D-CDC-T. In terms of the depth and optical flow modalities, the best performance (42.86% and 52%) could be achieved when using 3D-CDC-TR with θ=0.3 and θ=0.4, respectively. Compared with 3D-CDC-T, 3D-CDC-TR is more robust for depth and optical flow modalities because it alleviates sensor noises and pre-processing artifacts between frames in these two modalities. By the observation, the 3D-CDC-ST performs relatively poorly. The reason might be that the gesture recognition task prefers more temporal reasoning context than spatial gradient cues and appearance details. According to their enhanced temporal representation abilities, 3D-CDC-T with θ=0.6 and 3D-CDC-TR with θ=0.3 are considered in our NAS operation space.

Impact of Multi-Rate Branches. As gestures have various temporal scales, modeling such visual tempos of different gestures facilitates their recognition. Here we conduct the abla-

TABLE I: Comparison among various configurations of the two-stage NAS for varied modalities. The upper part is about the first stage NAS1 while bottom part is about the second stage NAS2. The evaluation metric is accuracy (%).

Configuration	RGB	Depth	RGB-D
w/o CDC, w/o NAS	44.07	43.10	-
w/o CDC, w/ NAS1	44.86	43.57	-
w/ CDC, w/o NAS	47.26	44.28	-
w/ CDC, w/ NAS1	48.38	45.08	-
w/ CDC, w/ NAS2_Fixed	-	-	42.62
w/ CDC, w/ NAS2_Shared	-	-	46.43
w/ CDC, w/ NAS2_Unshared	-	-	50.17

tion study with C3D [29] network to explore how the branches with different frame rates influence the gesture recognition task. As illustrated in Fig. 5, for the single rate network, the higher the frame rate it has, the better performance it will be. This is because a higher frame rate usually has less sampling temporal information loss, which has richer fine-grained temporal cues for gesture recognition. Furthermore, we could find that the performance could be further improved when cooperated with the multi-rate branches (e.g., '16 + 32 frames' and '8 + 16 + 32 frames'). In terms of the impact of multi-rate branches for different modalities, it is obvious that multi-rate branches contribute more to RGB than depth modality. When assembling with 3D-CDC-T for RGB or 3D-CDC-TR for depth, the trends of multi-rate networks are analogous as the vanilla cases but achieving holistic performance gains (due to the excellent representation capacity of 3D-CDC).

We also reimplement SlowFast [15] Networks (trained from scratch) with '8+32 frames' multi-rate setting on the IsoGD dataset. However, it only achieves respective 22.28% and 40.69% accuracy on RGB and depth inputs, which indicates the importance of suitable architecture design for multi-rate networks in the gesture recognition task.

Effectiveness of the Two-stage NAS. Based on the best multi-rate setting ('8+16+32 frames'), we study the two-stage NAS for both single and multiple modalities. The first stage NAS (**NAS1**) intends to find well-suited multi-rate single-modal networks. As illustrated in Tab. I, when searching over the vanilla search space w/o CDC, the architectures found by NAS1 improve 0.79% and 0.47% accuracy (compared with multi-rate C3D) for RGB and depth inputs, respectively. Moreover, the gains consistently occur when searching over 3D-CDC based search space for both RGB (+1.12%) and depth (+0.8%) modalities.

Based on the searched multi-rate networks for RGB and depth modalities, 'NAS2_Fixed' utilizes late fusion directly without searching the lateral connections between two modalities. Out of expectation, it performs even worse than the single-modal NAS1 searched models. It means that simply late fusion will encounter the problem of insufficient information exchange in feature levels. With the second stage NAS (**NAS2**), 'NAS2_Unshared' achieves more than 50% accuracy, which indicates the advantages of NAS that mines the efficient integration of multi-rate and multi-modal branches. Furthermore, compared with 'NAS2_Shared' searching the shared lateral connections for low-mid-high levels, 'NAS2_Unshared' performs better (+3.74%), which implies the importance of the specific design for interactions of each level.

TABLE II: Results on the validation set of IsoGD [2].

Method	Modality	Accuracy (%)
Hu *et al*.[66]	RGB	44.88
Res3D [8]	RGB	45.07
Zhang *et al*.[45]	RGB	51.31
Zhang *et al*.[11]	RGB	55.98
Zhu *et al*.[32]	RGB	57.42
NAS1 (Ours)	RGB	**58.88**
Narayana *et al*.[38]	Depth	27.98
Res3D [8]	Depth	48.44
Zhang *et al*.[45]	Depth	49.81
Zhu *et al*.[32]	Depth	54.18
NAS1 (Ours)	Depth	**55.68**
Zhu *et al*.[67]	RGB-D	51.02
Hu *et al*.[66]	RGB-D	54.14
Zhang *et al*.[45]	RGB-D	55.29
Zhu *et al*.[32]	RGB-D	61.05
NAS2 (Ours)	RGB-D	60.04
NAS1+NAS2 (Ours)	RGB-D	**65.54**
Li *et al*.[39]	RGB-D-Flow	54.50
Zhang *et al*.[45]	RGB-D-Flow	58.65
Miao *et al*.[7]	RGB-D-Flow	64.40
NAS2 (Ours)	RGB-Flow	61.22
NAS2 (Ours)	Flow-Depth	62.47
NAS2_all (Ours)	RGB-D-Flow	**66.23**
FOANet w/o hand[38]	RGB-D-Flow-DFlow	61.40
FOANet w/ hand [38]	RGB-D-Flow-DFlow	**80.96**

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

After studying the components in Sec. IV-C, we evaluate our models on three benchmark datasets. Note that in this subsection, our models are firstly pre-trained on the Jester [68] gesture dataset, which is similar to [45], [32], [69].

Results on IsoGD. As shown in Table II, although the existing methods [8], [45], [11] adopt 3DCNNs to learn from single RGB or depth modality, it is still challenging to represent the discriminative and robust spatio-temporal features with vanilla 3D convolutions and coarsely designed architecture. With the enhanced temporal representation capacity via 3D-CDC and multi-rate collaboration, our proposed multi-rate single-modal NAS method 'NAS1' obtains the best accuracy on every single modality. This exactly demonstrates the superiority of the searched architecture. In terms of the RGB-D gesture recognition, our searched architecture with two-stage NAS (NAS2) obtains more than 1% and 4% accuracy gains compared with the 'NAS1' with RGB and Depth modality, respectively. It demonstrates the effectiveness of RGB-D-temporal synergies at earlier stages. Similar to [32] ensembling the results from varied modalities, our 'NAS1+NAS2' boosts the accuracy to 65.54%.

To evaluate the modality generalization of the architecture searched from RGB-D, we retrain the same model 'NAS2' with RGB-Flow and Flow-Depth modalities separately and also achieve comparable performance (61.22% and 62.47%,

TABLE III: Results on the NVGesture [1] dataset.

Method	Modality	Accuracy (%)
HOG+HOG^2 [3]	RGB	24.50
Simonyan *et al.*[27]	RGB	54.60
Wang *et al.*[70]	RGB	59.10
C3D [29]	RGB	69.30
R3DCNN [1]	RGB	74.10
GPM [71]	RGB	75.90
PreRNN [33]	RGB	76.50
ResNeXt-101 [69]	RGB	78.63
MTUT [10]	RGB	81.33
NAS1 (Ours)	RGB	**83.61**
HOG+HOG^2 [3]	Depth	36.30
SNV [72]	Depth	70.70
C3D [29]	Depth	78.80
R3DCNN [1]	Depth	80.30
ResNeXt-101 [69]	Depth	83.82
PreRNN [33]	Depth	84.40
MTUT [10]	Depth	84.85
GPM [71]	Depth	85.50
NAS1 (Ours)	Depth	**86.10**
HOG+HOG^2 [3]	RGB-D	36.90
I3D [31]	RGB-D	83.82
PreRNN [33]	RGB-D	85.00
MTUT [10]	RGB-D	85.48
GPM [71]	RGB-D	86.10
NAS2 (Ours)	RGB-D	86.93
NAS1+NAS2 (Ours)	RGB-D	**88.38**

respectively). To our best knowledge, it is the first to explore the modality generalization issues for multi-modal NAS. Finally, the best accuracy (66.23%) could be achieved when ensembling the scores from all three 'NAS2' models among RGB-D-Flow modalities. Although the FOANet [38] reports the best performance (80.96%) on IsoGD, the high accuracy is achieved by fusing 12 channels (i.e., global/left/right channels for four modalities) with manual hand detection. Note that without hand detection preprocessing, our 'NAS2_all' outperforms FOANet (66.23% vs. 61.4%) by a large margin using only RGB-D-Flow modalities. This exactly demonstrates the superiority of our searched architectures.

Results on NVGesture. Table III compares the performance of our method with SOTA methods on the NVGesture dataset. It can be seen that our approach performs the best for both single-modal and multi-modal testing, which indicates 1) our searched architecture is able to represent discriminative spatio-temporal features for single/multi-modal gesture recognition; and 2) the architecture searched on the source dataset (IsoGD) via the proposed two-stage NAS transfers well on the target dataset (NVGesture), demonstrating the excellent generalization ability of the proposed NAS method.

Fig. 6 evaluates the coherence between the predicted labels from the searched 'NAS1' and 'NAS2' architectures, and the ground truths on the NVGesture dataset. The coherence is calculated by their confusion matrices. We observe that with RGB-D-temporal synergies, 'NAS2' has less confusion between the input classes and provides generally a more diagonalized confusion matrix. This improvement is better observed in the first three and last six classes.

Results on EgoGesture. We also evaluate the robustness of our searched architectures on egocentric (first-person view) gesture recognition. Compared with the the third-person view gesture recognition, the main challenges include the complex scene background and motion blurriness caused by the subject walking. It can be seen in Table IV that our approach achieves the best performance ('NAS2' with 94.38% and 'NAS1+NAS2' with 95.52%) on the EgoGesture dataset using RGB-D modalities, which indicates the strong generalization ability of the proposed 3D-CDC and two-stage NAS method. Note that ResNeXt-101 [69] needs an extra detector to capture the key segments for preprocessing. As our proposed multi-rate and multi-modal network recognizes the gesture on original video clips directly, the performance might be further boosted with the gesture detector.

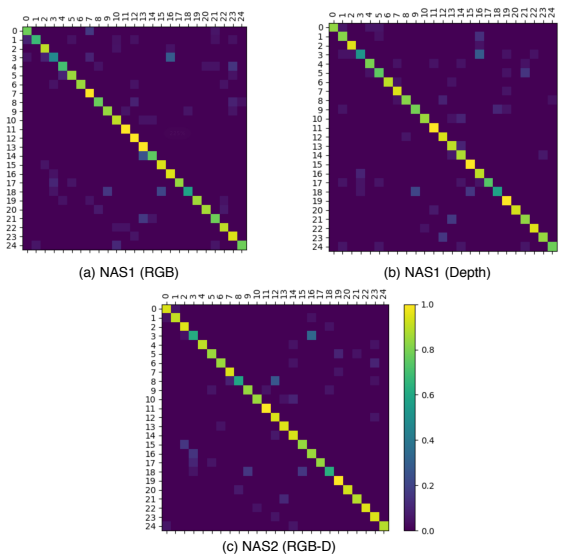


Fig. 6: The confusion matrices obtained by comparing the ground-truth labels and the predicted labels from the NAS1 and NAS2 networks trained on the NVGesture dataset. Best seen on the computer, in color and zoomed in.

TABLE IV: Results on the EgoGesture [73] dataset.

Method	Modality	Accuracy (%)
VGG16+LSTM [27]	RGB	74.7
C3D+LSTM+RSTTM [29]	RGB	89.3
ResNeXt-101 [69]	RGB	**93.75**
NAS1 (Ours)	RGB	93.31
VGG16+LSTM [72]	Depth	77.7
C3D+LSTM+RSTTM [1]	Depth	90.6
ResNeXt-101 [69]	Depth	94.03
NAS1 (Ours)	Depth	**94.13**
VGG16+LSTM [31]	RGB-D	81.4
C3D+LSTM+RSTTM [10]	RGB-D	92.2
I3D [71]	RGB-D	92.78
MTUT^F [71]	RGB-D	93.87
NAS2 (Ours)	RGB-D	94.38
NAS1+NAS2 (Ours)	RGB-D	**95.52**

V. Discussion and Visualization

In this section, we first discuss the transferability of the proposed two-stage NAS and 3D-CDC on action recognition

TABLE V: Results on the RGB-D action recognition dataset THU-READ [73] with the cross-subject protocol. The architectures of 'NAS1' and 'NAS2' are searched on IsoGD and then retrained/evaluated on THU-READ.

Method	Modality	Accuracy(%)
HOG [74]	RGB	39.93
HOF [75]	RGB	46.27
Appearance Stream [76]	RGB	41.90
C3D [29]	RGB	66.25
SlowFast [15]	RGB	69.58
NAS1 (Ours)	RGB	**71.25**
HOG [74]	Depth	45.83
HOF [75]	Depth	43.96
Depth Stream [76]	Depth	34.06
C3D [29]	Depth	63.75
SlowFast [15]	Depth	68.75
NAS1 (Ours)	Depth	**69.58**
MDNN[73]	RGB-D-Flow	62.92
C3D [29]	RGB-D	75.83
SlowFast [15]	RGB-D	76.25
NAS2 (Ours)	RGB-D	73.85
NAS1+NAS2 (Ours)	RGB-D	**78.38**

task, which is interesting and necessary because it exists task-oriented biases (gesture recognition is less relied on the scene but more related to the fine-grained temporal cues when compared with the action recognition). Then, we give the visualization of the searched architecture and feature response.

A. Task Transferability

Generalization to RGB-D Action Recognition. In order to validate the generalization ability of our 3D-CDC based two-stage NAS, we transfer the searched architecture ('NAS1' and 'NAS2') to another multi-modal video understanding task, i.e., RGB-D action recognition. Here one of the largest RGB-D egocentric dataset THU-READ [73] is used for experiments. It consists of 40 different actions and 1920 videos. We adopt the released leave-one-split-out protocol. For fair comparison, C3D [29], SlowFast [15], 'NAS1', and 'NAS2' are pre-trained on Jester gesture dataset. Table V compares the performance of our method with SOTA methods on THU-READ. It can be seen that our approach outperforms the mainstream 3DCNNs (e.g., C3D [29] and SlowFast [15]) with a convincing margin, indicating that the architecture searched on the source task (gesture recognition) could be generalized well on the target video understanding task (e.g., action recognition).

Impact of 3D-CDC for RGB Action Recognition. Here we explore the effectiveness of 3D-CDC for scene-based action recognition tasks. Fig. 7 illustrates the results of two classical scene-related action datasets (UCF101 [77] and HMDB51 [78]). It is obvious that compared with the 3D vanilla convolution ($\theta = 0$), 3D-CDC-ST improves the performance dramatically in both datasets (+3% for UCF101 when $\theta = 0.6$ and +2.1% for HMDB51 when $\theta = 0.4, 0.6, 0.8$). The reason might be twofold. On one hand, an enhanced spatio-temporal context is helpful to represent scene-aware appearance and motions. On the other hand, the spatio-temporal difference term can be regarded as a regularization

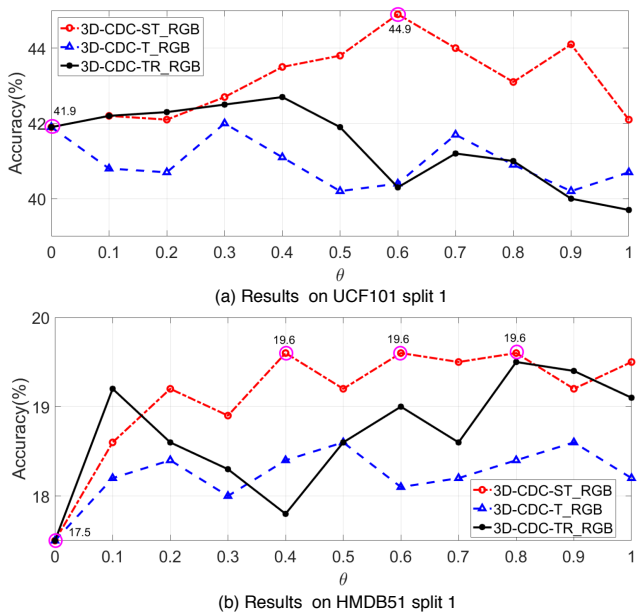


Fig. 7: Impacts of the 3D-CDC family on two benchmark action recognition datasets (a) UCF101 [77], and (b) HMDB51 [78]. We use 3D-ResNet18 [30] as the backbone and sequence size $3\times16\times112\times112$ as the inputs. All experiments are trained from scratch for fair comparison.

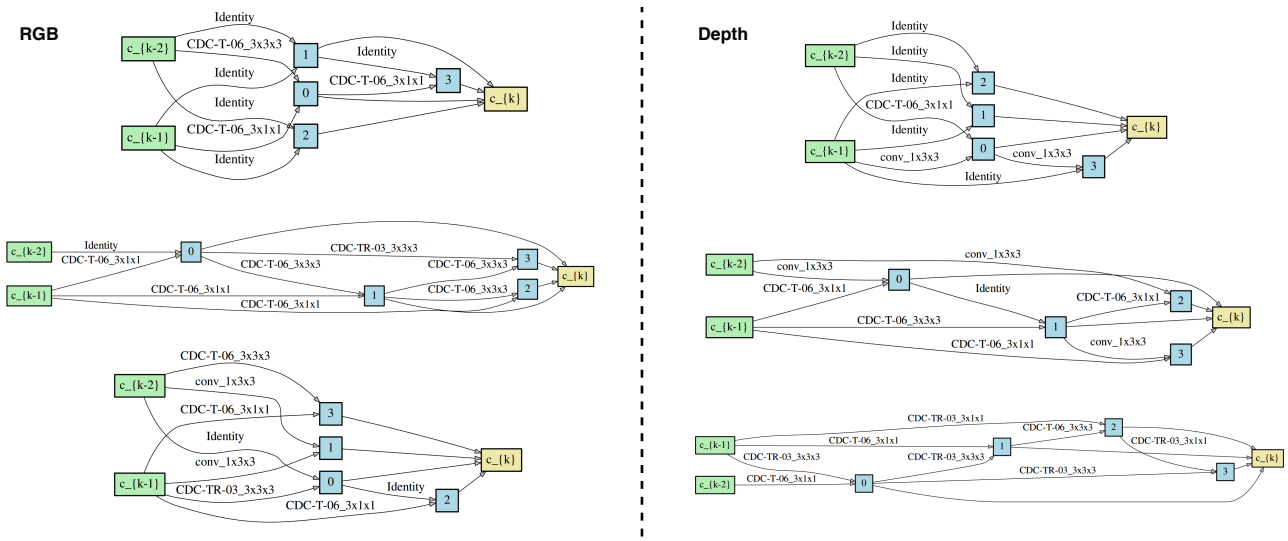
term to alleviate overfitting. Another highlight is that, without extra parameters, 3D-ResNet18 assembled with 3D-CDC-ST outperforms that with Spatio-Temporal Channel Correlation (STC) Block [34] by +2.1% on UCF101 split 1. In contrast, 3D-CDC-T performs the worst because of its weak spatial context representation capacity and vulnerability to the scene noises, which are vital in these two scene-aware datasets.

B. Visualization of the Searched Architecture

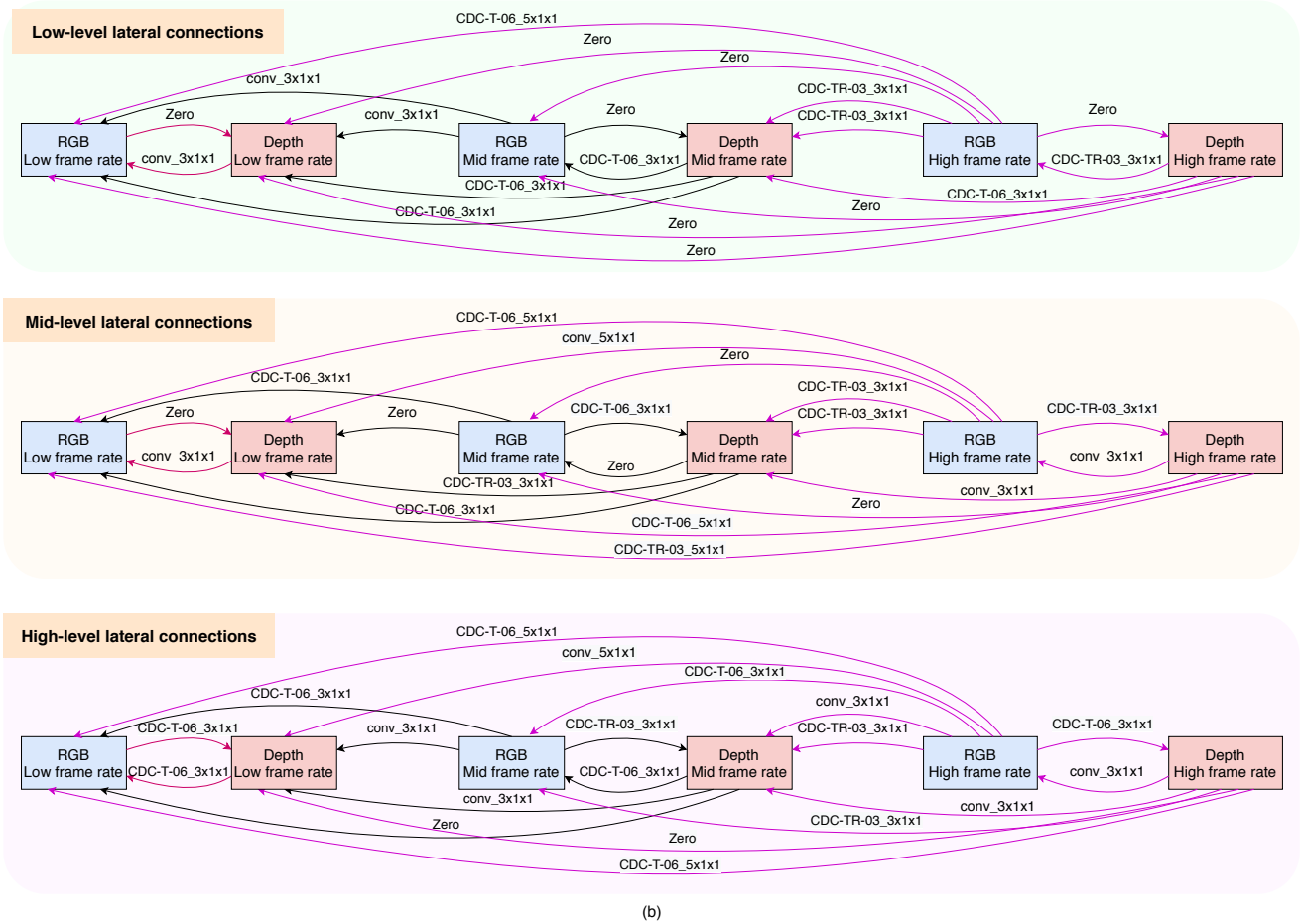
Fig. 8 shows the searched cells and lateral connections with the proposed two-stage NAS. It can be seen from Fig. 8(a) that there are more 'CDC-T-06' based operators in all three RGB branches while more 'CDC-TR-03' based operators in the depth branches. This consists with the observations in our ablation study of 'Impact of 3D-CDC for Modalities' in Section 4.1. In Fig. 8(b), it is interesting to find that the lower-level lateral connections are sparser (i.e., with more 'Zero' operators) and the high-level lateral connections have more learnable operators (i.e., convolution operators). This might inspire the video understanding community to design more reasonable multi-modal networks in the future.

C. Feature Visualization

The neural activation (before Pool3 in C3D) are visualized in Fig. 9. It can be seen from Fig. 9(a) that the proposed 3D-CDC-ST, 3D-CDC-T, and 3D-CDC-TR enhance the spatio-temporal representation and enforce the model to focus more on the trajectories of arms and hands. As for the depth modality shown in Fig. 9(b), all the convolutions are able to make the accurate attention on the movements from arms and hands due to the benefits from the foregrounds provided by



(a)



(b)

Fig. 8: The searched architecture from (a) the first stage NAS, and (b) the second stage NAS. The three rows in (a) represent the searched cell structure in the low, mid, and high frame branches, respectively.

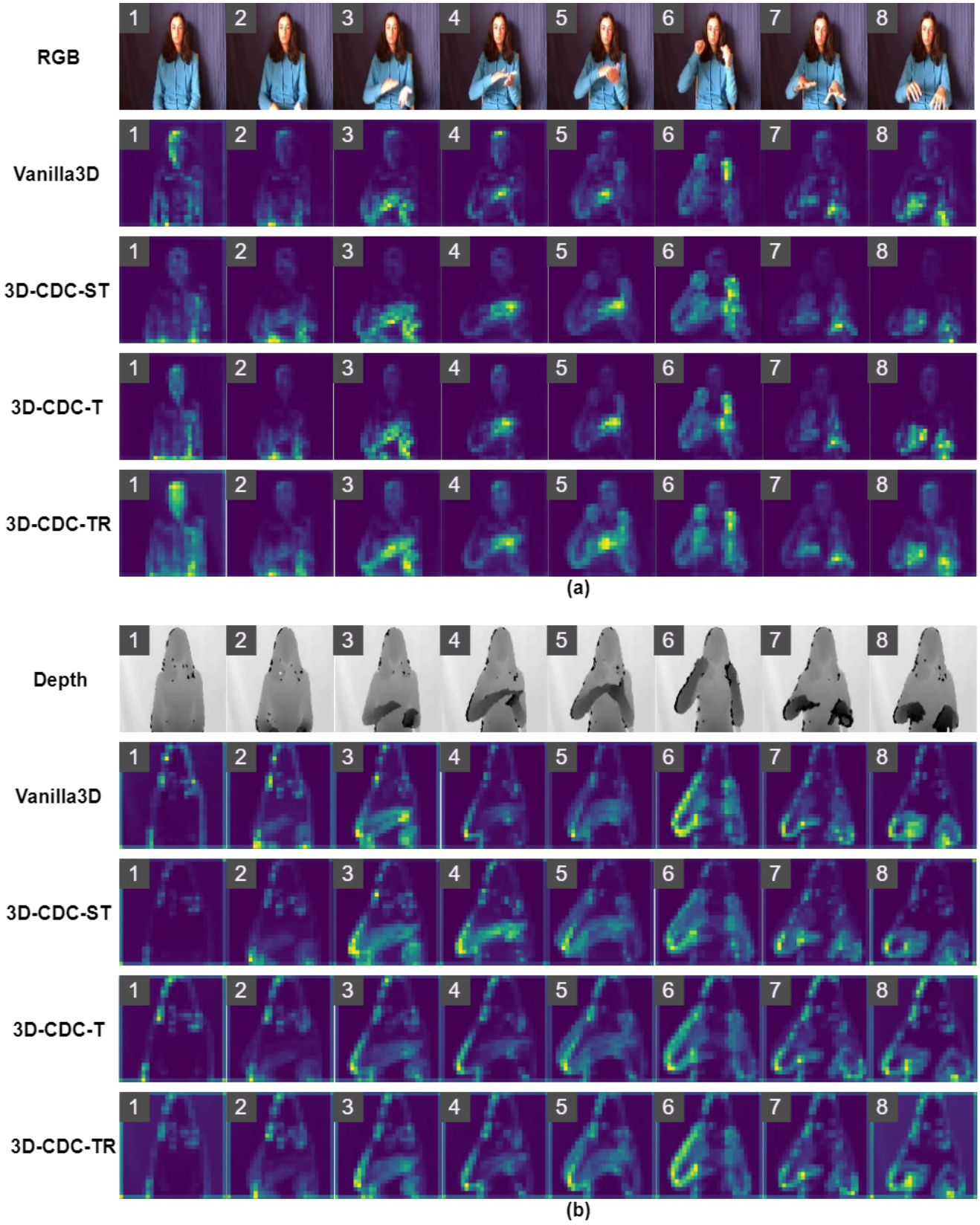


Fig. 9: Features visualization from C3D assembled with varied convolutions on the IsoGD dataset. With (a) RGB and (b) Depth modality inputs, the four rows represent the neural activation with 3D vanilla convolution, 3D-CDC-ST, 3D-CDC-T, and 3D-CDC-TR, respectively. Best view when zoom in.

the depth inputs. Despite more robust representation achieved by the 3D-CDC family, the interferences from the sensor-based noise and undesirable movements (e.g., head and hair) still occur. Thus, it is necessary to explore RGB-D-temporal synergies to overcome such limitation.

VI. Conclusion

We present a novel 3D convolution family called 3D-CDC for enhancing the spatio-temporal representation for video understanding tasks. Over the 3D-CDC search space, we propose a two-stage NAS method to discover well-suited multi-rate and multi-modal networks with RGB-D-temporal synergies. Extensive experiments show the effectiveness of our method. Future directions include: 1) exploring 3D-CDC family on other video understanding tasks (e.g., temporal localization); 2) searching temporal synergies with more modalities (e.g., audio and skeleton).

VII. Acknowledgment

This work was supported by the Academy of Finland for project MiGA (grant 316765), ICT 2023 project (grant 328115), Infotech Oulu, and the Chinese National Natural Science Foundation Projects #61961160704, #61876179, Science and Technology Development Fund of Macau No. 0025/2019/A1. The authors also wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

References

[1] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network," in *CVPR*, 2016, pp. 4207–4215.

[2] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li, "Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition," in *CVPR Workshops*, 2016, pp. 56–64.

[3] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *TITS*, vol. 15, no. 6, pp. 2368–2377, 2014.

[4] Y. Zhang, C. Cao, J. Cheng, and H. Lu, "Egogesture: a new dataset and benchmark for egocentric hand gesture recognition," *TMM*, vol. 20, no. 5, pp. 1038–1050, 2018.

[5] J. Weissmann and R. Salomon, "Gesture recognition for virtual reality applications using data gloves and neural networks," in *IJCNN*, vol. 3. IEEE, 1999, pp. 2043–2046.

[6] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial intelligence review*, vol. 43, no. 1, pp. 1–54, 2015.

[7] Q. Miao, Y. Li, W. Ouyang, Z. Ma, X. Xu, W. Shi, and X. Cao, "Multimodal gesture recognition based on the resc3d network," in *ICCV Workshops*, 2017, pp. 3047–3055.

[8] Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, R. Li, and J. Song, "Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model," in *ICPR*. IEEE, 2016, pp. 25–30.

[9] H. Wang, P. Wang, Z. Song, and W. Li, "Large-scale multimodal gesture segmentation and recognition based on convolutional neural networks," in *ICCV*, 2017, pp. 3138–3146.

[10] M. Abavisani, H. R. V. Joze, and V. M. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," in *CVPR*, 2019, pp. 1165–1174.

[11] L. Zhang, G. Zhu, L. Mei, P. Shen, S. A. A. Shah, and M. Bennamoun, "Attention in convolutional lstm for gesture recognition," in *NeurIPS*, 2018, pp. 1953–1962.

[12] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *CVPR*, 2018, pp. 1390–1399.

[13] M. Lee, S. Lee, S. Son, G. Park, and N. Kwak, "Motion feature network: Fixed motion filter for action recognition," in *ECCV*, 2018, pp. 387–403.

[14] A. Piergiovanni and M. S. Ryoo, "Representation flow for action recognition," in *CVPR*, 2019, pp. 9945–9953.

[15] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *ICCV*, 2019, pp. 6202–6211.

[16] D. Zhang, X. Dai, and Y.-F. Wang, "Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection," in *ACCV*. Springer, 2018, pp. 712–728.

[17] Y. Wang, M. Long, J. Wang, and P. S. Yu, "Spatiotemporal pyramid network for video action recognition," in *CVPR*, 2017, pp. 1529–1538.

[18] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: adaptive multi-modal gesture recognition," *TPAMI*, vol. 38, no. 8, pp. 1692–1706, 2015.

[19] V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos, "Multi-modal gesture recognition via multiple hypotheses rescoring," in *Gesture Recognition*. Springer, 2017, pp. 467–496.

[20] A. Roitberg, T. Pollert, M. Haurilet, M. Martin, and R. Stiefelhagen, "Analysis of deep fusion strategies for multi-modal gesture recognition," in *CVPR Workshops*, 2019, pp. 0–0.

[21] J. Wan, Q. Ruan, W. Li, G. An, and R. Zhao, "3d smosift: three-dimensional sparse motion scale invariant feature transform for activity recognition from rgb-d videos," *JEI*, vol. 23, no. 2, p. 023017, 2014.

[22] J. Wan, Q. Ruan, W. Li, and S. Deng, "One-shot learning gesture recognition from rgb-d data using bag of features," *JMLR*, vol. 14, no. 1, pp. 2549–2582, 2013.

[23] M. R. Malgireddy, I. Inwogu, and V. Govindaraju, "A temporal bayesian model for classifying, detecting and localizing activities in video sequences," in *CVPR Workshops*. IEEE, 2012, pp. 43–48.

[24] J. Wan, G. Guo, and S. Z. Li, "Explore efficient local features from rgb-d data for one-shot learning gesture recognition," *TPAMI*, vol. 38, no. 8, pp. 1626–1639, 2015.

[25] X. Ji, J. Cheng, D. Tao, X. Wu, and W. Feng, "The spatial laplacian and temporal energy pyramid representation for human action recognition using depth sequences," *Knowledge-Based Systems*, vol. 122, pp. 64–74, 2017.

[26] Z. Liu, X. Chai, Z. Liu, and X. Chen, "Continuous gesture recognition with hand-oriented spatiotemporal feature," in *ICCV Workshops*, 2017, pp. 3056–3064.

[27] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014, pp. 568–576.

[28] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, "Cooperative training of deep aggregation networks for rgb-d action recognition," in *AAAI*, 2018.

[29] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[30] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *CVPR*, 2018, pp. 6546–6555.

[31] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[32] G. Zhu, L. Zhang, L. Yang, L. Mei, S. A. A. Shah, M. Bennamoun, and P. Shen, "Redundancy and attention in convolutional lstm for gesture recognition," *TNNLS*, 2019.

[33] X. Yang, P. Molchanov, and J. Kautz, "Making convolutional networks recurrent for visual sequence learning," in *CVPR*, 2018, pp. 6469–6478.

[34] A. Diba, M. Fayyaz, V. Sharma, M. Mahdi Arzani, R. Yousefzadeh, J. Gall, and L. Van Gool, "Spatio-temporal channel correlation networks for action classification," in *ECCV*, 2018, pp. 284–299.

[35] S. Kumawat and S. Raman, "Lp-3dcnn: Unveiling local phase in 3d convolutional neural networks," in *CVPR*, 2019, pp. 4903–4912.

[36] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," *CVPR*, 2020.

[37] Z. Yu, Y. Qin, X. Li, Z. Wang, C. Zhao, Z. Lei, and G. Zhao, "Multi-modal face anti-spoofing based on central difference networks," in *CVPR Workshops*, 2020, pp. 650–651.

[38] P. Narayana, R. Beveridge, and B. A. Draper, "Gesture recognition: Focus on the hands," in *CVPR*, 2018, pp. 5235–5244.

[39] Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, Z. Ma, and J. Song, "Large-scale gesture recognition with a fusion of rgb-d data based on optical flow and the c3d model," *PR Letters*, vol. 119, pp. 187–194, 2019.

[40] X. Liu, H. Shi, X. Hong, H. Chen, D. Tao, and G. Zhao, "3d skeletal gesture recognition via hidden states exploration," *TIP*, vol. 29, pp. 4583–4597, 2020.

[41] X. Liu and G. Zhao, "3d skeletal gesture recognition via discriminative coding on time-warping invariant riemannian trajectories," *TMM*, 2020.

[42] Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, R. Li, and J. Song, "Large-scale gesture recognition with a fusion of rgb-d data based on saliency theory and c3d model," *TCSVT*, vol. 28, no. 10, pp. 2956–2964, 2017.

[43] J. Duan, J. Wan, S. Zhou, X. Guo, and S. Z. Li, "A unified framework for multi-modal isolated gesture recognition," *TOMM*, vol. 14, no. 1s, pp. 1–16, 2018.

[44] G. Zhu, L. Zhang, L. Mei, J. Shao, J. Song, and P. Shen, "Large-scale isolated gesture recognition using pyramidal 3d convolutional networks," in *ICPR*. IEEE, 2016, pp. 19–24.

[45] L. Zhang, G. Zhu, P. Shen, J. Song, S. Afaq Shah, and M. Bennamoun, "Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition," in *ICCV Workshops*, 2017, pp. 3120–3128.

[46] X. Dong and Y. Yang, "Searching for a robust neural architecture in four gpu hours," in *CVPR*, 2019, pp. 1761–1770.

[47] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," *ICLR*, 2019.

[48] H. Pham, M. Y. Guan, B. Zoph, Q. V. Le, and J. Dean, "Efficient neural architecture search via parameter sharing," *ICML*, 2018.

[49] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *ICLR*, 2017.

[50] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *CVPR*, 2018, pp. 8697–8710.

[51] Y. Xu, L. Xie, X. Zhang, X. Chen, G.-J. Qi, Q. Tian, and H. Xiong, "Pc-darts: Partial channel connections for memory-efficient architecture search," in *ICLR*, 2019.

[52] Z. Yu, Y. Qin, X. Xu, C. Zhao, Z. Wang, Z. Lei, and G. Zhao, "Auto-fas: Searching lightweight networks for face anti-spoofing," in *ICASSP*. IEEE, 2020, pp. 996–1000.

[53] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *AAAI*, vol. 33, 2019, pp. 4780–4789.

[54] E. Real, S. Moore, A. Selle, S. Saxena, Y. L. Suematsu, J. Tan, Q. V. Le, and A. Kurakin, "Large-scale evolution of image classifiers," in *ICML*, 2017, pp. 2902–2911.

[55] H. Cai, L. Zhu, and S. Han, "Proxylessnas: Direct neural architecture search on target task and hardware," *ICLR*, 2019.

[56] W. Peng, X. Hong, and G. Zhao, "Video action recognition via neural architecture searching," in *ICIP*. IEEE, 2019, pp. 11–15.

[57] Z. Qiu, T. Yao, Y. Zhang, Y. Zhang, and T. Mei, "Scheduled differentiable architecture search for visual recognition," *arXiv preprint arXiv:1909.10236*, 2019.

[58] X. Wang, X. Xiong, M. Neumann, A. Piergiovanni, M. S. Ryoo, A. Angelova, K. M. Kitani, and W. Hua, "Attentionnas: Spatiotemporal attention cell search for video classification," *arXiv preprint arXiv:2007.12034*, 2020.

[59] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "Mfas: Multimodal fusion architecture search," in *CVPR*, 2019, pp. 6966–6975.

[60] M. S. Ryoo, A. Piergiovanni, M. Tan, and A. Angelova, "Assemblenet: Searching for multi-stream neural connectivity in video architectures," *ICLR*, 2020.

[61] M. Tan and Q. V. Le, "Mixconv: Mixed depthwise convolutional kernels," *arXiv preprint arXiv:1907.09595*, 2019.

[62] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *ICCV*, 2019, pp. 1314–1324.

[63] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV*. Springer, 2004, pp. 25–36.

[64] J. Wan, C. Lin, L. Wen, Y. Li, Q. Miao, S. Escalera, G. Anbarjafari, I. Guyon, G. Guo, and S. Z. Li, "Chalearn looking at people: Isogd and congd large-scale rgb-d gesture recognition," *arXiv preprint arXiv:1907.12193*, 2019.

[65] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan, "Learning features by watching objects move," in *CVPR*, 2017, pp. 2701–2710.

[66] T.-K. Hu, Y.-Y. Lin, and P.-C. Hsiu, "Learning adaptive hidden layers for mobile gesture recognition," in *AAAI*, 2018.

[67] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal gesture recognition using 3-d convolution and convolutional lstm," *Ieee Access*, vol. 5, pp. 4517–4524, 2017.

[68] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *ICCV Workshops*, 2019, pp. 0–0.

[69] O. Köpüklü, A. Gunduz, N. Kose, and G. Rigoll, "Real-time hand gesture detection and classification using convolutional neural networks," in *FG*. IEEE, 2019, pp. 1–8.

[70] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *IJCV*, vol. 119, no. 3, pp. 219–238, 2016.

[71] V. Gupta, S. K. Dwivedi, R. Dabral, and A. Jain, "Progression modelling for online and early gesture detection," in *3DV*. IEEE, 2019, pp. 289–297.

[72] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *CVPR*, 2014, pp. 804–811.

[73] Y. Tang, Z. Wang, J. Lu, J. Feng, and J. Zhou, "Multi-stream deep neural networks for rgb-d egocentric action recognition," *TCSVT*, vol. 29, no. 10, pp. 3001–3015, 2018.

[74] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," 2009.

[75] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*. IEEE, 2008, pp. 1–8.

[76] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[77] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[78] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*. IEEE, 2011, pp. 2556–2563.