Convolutional Memory Blocks for Depth Data Representation Learning

Keze Wang[1,2], **Liang Lin**[1,3*], **Chuangjie Ren**[1], **Wei Zhang**[3], and **Wenxiu Sun**[3]
[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] The Hong Kong Polytechnic University
[3] Sensetime Group Limited
{kezewang,chuangjieren,irene.wenxiu.sun}@gmail.com,
linliang@ieee.org, wayne.zhang@sensetime.com

Abstract

Compared to natural RGB images, data captured by 3D / depth sensors (e.g., Microsoft Kinect) have different properties, e.g., less discriminable in appearance due to lacking color / texture information. Applying convolutional neural networks (CNNs) on these depth data would lead to unsatisfying learning efficiency, i.e., requiring large amounts of annotated training data for convergence. To address this issue, this paper proposes a novel memory network module, called Convolutional Memory Block (CMB), which empowers CNNs with the memory mechanism on handling depth data. Different from the existing memory networks that store long / short term dependency from sequential data, our proposed CMB focuses on modeling the representative dependency (correlation) among non-sequential samples. Specifically, our CMB consists of one internal memory (i.e., a set of feature maps) and three specific controllers, which enable a powerful yet efficient memory manipulation mechanism. In this way, the internal memory, being implicitly aggregated from all previous inputted samples, can learn to store and utilize representative features among the samples. Furthermore, we employ our CMB to develop a concise framework for predicting articulated pose from still depth images. Comprehensive evaluations on three public benchmarks demonstrate significant superiority of our framework over all the compared methods. More importantly, thanks to the enhanced learning efficiency, our framework can still achieve satisfying results using much less training data.

1 Introduction

With the rapid development of inexpensive commodity depth sensors, depth data representation learning is ubiquitous in many applications such as robotic systems [McColl *et al.*, 2011]. Compared to RGB data which provides information about appearance and texture, depth data, reflecting the distance information between the objects and the sensor, are less discriminable. Recently, deep convolutional neural networks (CNNs) have been applied by many methods [Haque *et al.*, 2016; Wang *et al.*, 2016] for depth data analysis. Due to the heavy sensor noise of depth data and the huge parameters of used deep CNNs, these methods require plenty of well annotated samples to train from scratch to achieve the satisfactory performance. However, collecting and annotating on depth data is extremely laborious and time-consuming. It will be beneficial to design a more intelligent network architecture with better learning efficiency on depth data, i.e., to surpass the state-of-the-art performance even with less training data.

In this paper, we propose a novel memory network module,

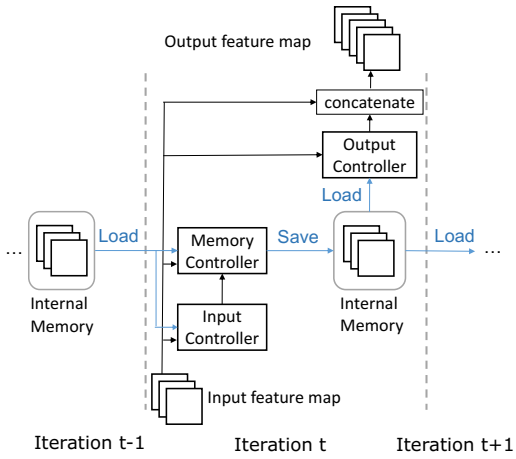

Figure 1: Detailed architecture of our proposed Convolutional Memory Block (CMB). The CMB consists of one internal memory (i.e., a set of feature maps) and three specific convolutional controllers, which can manipulate the internal memory and extract implicit structural representation from the input feature map. Specifically, the old internal memory from the previous training iteration is loaded by the input and memory controller. Then, the memory controller fuses its memory representation and the response of the input controller, and saves the fused representation to be the new internal memory. After that, the output controller loads the new internal memory to generate memory representation, which is further concatenated with the input feature map to be the final output.

*Corresponding author is Liang Lin (Email: linliang@ieee.org).

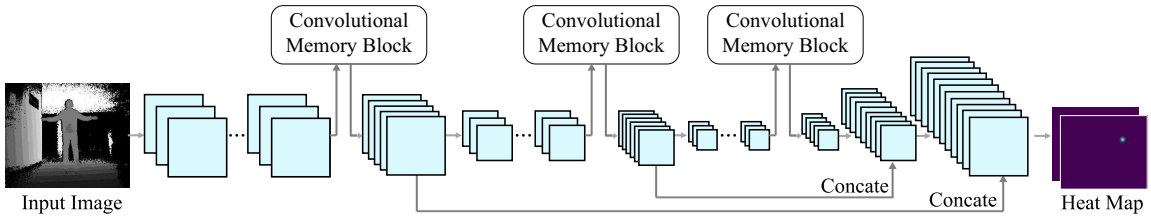

Figure 2: An overview of the proposed convolutional memory block embedded framework for estimating articulated poses.

called Convolutional Memory Block (CMB), inspired by the recent work of Neural Turing Machine [Graves *et al.*, 2014]. Considering the special properties of depth data, e.g., low discriminability in appearance due to lacking color / texture information, we leverage the memory mechanism to enhance the pattern abstracting of CNNs by reusing their rich implicit convolutional structures and spatial correlations among the training samples. Specifically, the proposed CMB consists of three specific convolutional controllers and one internal memory (i.e., a set of feature maps) as shown in Figure 1. The convolutional controller is designed to process the input feature map from the previous layer and manipulate the internal memory. Different from ConvLSTM [Shi *et al.*, 2015] and ConvGRU [Ballas *et al.*, 2016] that require time-series data, the proposed convolutional controller performs convolution in a hierarchical organization via several convolutional layers with batch normalization. This ensures that our proposed CMB is capable of extracting more abstract information from non-sequential training samples to augment image-dependent feature representation. Specifically, our CMB intends to capture and store the representative dependencies or correlations among training samples according to specific learning tasks, and further employ these stored dependencies to enhance the representation of convolutional layers. In this way, our CMB encourages the CNN architecture to be lightweight and require less training data.

Since articulated (e.g., human body or hand) pose estimation is one of the most dominant applications in depth data representation learning, we develop a simple yet effective articulated pose estimation framework to validate the effectiveness of our CMB by applying it to enhance the convolutional layers. Recently, highly accurate and real-time performance on human pose estimation from depth data has been achieved by deep learning based methods [Haque *et al.*, 2016; Wang *et al.*, 2016]. Nevertheless, all methods borrow CNN architectures from RGB data analysis. None of them considers how to improve the learning efficiency in the perspective of handling depth images. Motivated by the design principles from stacked hourglass [Newell *et al.*, 2016], our developed framework has a lightweight hourglass-like architecture, as illustrated in Figure 2. Our CMB contributes to collaboratively regress the heat map of each joint by providing the cached representative feature maps, which are aggregated from previous inputted samples.

The **main contributions** of this work are summarized as follows: (i) we propose a novel Convolutional Memory Block (CMB) to promote the representation and learning efficiency of convolutional layers for handling depth data; (ii) we apply our CMB to develop a simple yet effective framework for articulated pose estimation from depth images. Extensive experiments on three public benchmarks not only demonstrate the superiority of our framework over all the compared methods, but also prove our framework can obtain comparable performance with much less training data.

2 Related Work

Neural Networks with Memory. To model the temporal dynamics and dependency, Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] and Gated Recurrent Unit (GRU) [Cho *et al.*, 2014], have been proposed and achieved the remarkable performance on many vision tasks (e.g., semantic parsing [Liang *et al.*, 2016]). More recently, the Convolutional Long Short-Term Memory model (ConvLSTM) [Shi *et al.*, 2015] and Convolutional Gated Recurrent Unit (ConvGRU) [Ballas *et al.*, 2016] have been proposed to consider the correlations among neighboring pixels in the spatial domain by enabling convolutional operations among its gates, hidden states and memory cells.

Besides above-mentioned models, various memory mechanisms have been proposed to enable neural networks to model sequential data by explicitly remembering variables and data over long timescales. The existing memory network models can be divided into content-based addressing and location-based addressing according to their accessing memory manners. The location-based addressing (e.g., [Graves *et al.*, 2014]) leverages a module so-called controller to receive input data and further store or retrieve valuable information from an external memory via the looked up address, while the content-based addressing (e.g., [Graves *et al.*, 2016]) focuses on reading from or writing to the memory to obtain the relevant memory cells or representations instead. Specifically, Neural Turing Machines [Graves *et al.*, 2014] was first proposed to use an external memory to solve some algorithmic problems via location addressing. [Na *et al.*, 2017] proposed to design the read network and the write network that consist of multiple convolutional layers.

Articulated Pose Estimation from Depth Data. Recently, several works [Shotton *et al.*, 2011; 2013; Jung *et al.*, 2015] have achieved promising performances on articulated pose estimation such as human and hand pose estimation. [Shotton *et al.*, 2011] designed an intermediate body parts representation that maps the difficult pose estimation problem into a simpler per-pixel classification problem. [Jung *et al.*, 2015] proposed to introduce a regression tree to estimate human poses by applying a supervised gradient descent and

MCMC like random sampler in the form of Markov random walks. More recently, improved performances have also been achieved by deep learning based methods [Haque *et al.*, 2016; Wang *et al.*, 2016]. Specifically, [Haque *et al.*, 2016] presented a viewpoint invariant model by combining CNN and RNN with a top-down error feedback mechanism to self-correct previous pose estimates in an end-to-end manner. [Wang *et al.*, 2016] proposed an inference embedded multi-task learning framework, which is implemented with a deep architecture of neural networks with two cascaded tasks.

The advanced network architectures [Wei *et al.*, 2016; Newell *et al.*, 2016; Grinciunaite *et al.*, 2016] for estimating human poses from RGB images have also been proposed. [Wei *et al.*, 2016] provided a sequential human pose prediction framework for learning rich implicit spatial models. [Newell *et al.*, 2016] proposed to design network architecture by processing features across all scales and consolidate them to best capture the various spatial relationships associated with the body. However, directly applying these architectures to the depth data is unfeasible due to the different challenges and requirements between the RGB data and the depth data. For instance, the depth images usually include heavy sensor noises and preserve coarse appearance details.

3 Convolutional Memory Blocks

As illustrated in Figure 1, our proposed Convolutional Memory Block (CMB) consists of one internal memory (a set of feature maps) and three convolutional controllers for facilitating the internal memory manipulation. The internal memory is designed to store the learned implicit image-dependent structural features, and is denoted as $M \in \mathbb{R}^{c_m \times h_m \times w_m}$, where c_m, h_m, w_m are the capacity, height, width of the feature map, respectively. Different from the widely used neural controllers [Weston *et al.*, 2015; Park *et al.*, 2017] which use full connections to perform input-to-state and state-to-state transitions, our proposed convolutional controller leverages the convolution operator to process the input feature map from the previous neural layer, and further manipulates the internal memory in both the training and testing phase. Figure 1 also illustrates three specific convolutional controllers in our proposed CMB, i.e., input controller, memory controller and output controller.

Inspired by the design principles from [Szegedy *et al.*, 2015], our proposed convolutional controllers share the same elaborately crafted structure, which leverages a hierarchical organization of convolutional layers rather than a simple convolutional layer as ConvLSTM [Shi *et al.*, 2015] and ConvGRU [Ballas *et al.*, 2016]. This ensures that the convolutional controller is able to extract rich and high-level implicit structural features. As illustrated in Figure 3, each controller first concatenates both the input feature map and internal memory, and further employs a single 3×3 convolutional layer, Batch Normalization (BN) layer, Rectified Linear Units (ReLU), 1×1 convolutional layer and another BN layer to obtain the intermediate feature map. For ease of implementation, the channel number and stride of all the convolutional layers are the same. As reported by [Zhang *et al.*, 2018], the 1×1 convolutional layer is imposed to help the information flow across different channels of the feature map. For the convenience of the formulation, we simply denote the operations with the convolutional controller as the function $\phi(\cdot)$. Note that, to simplify the further calculation, the output response of each convolutional kernels all share the same size with the input feature map. For each specific convolution controller, the intermediate feature map is further processed individually. In the following, we will introduce these three specific convolutional controllers in the training and testing phase formally.

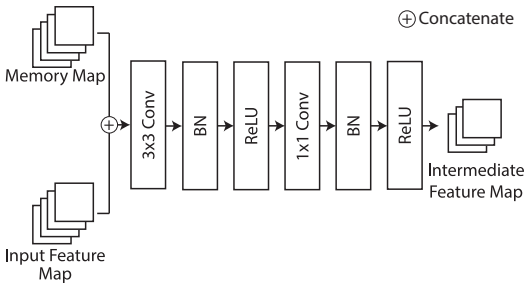

Figure 3: Detailed illustration of our convolutional controller.

3.1 Input Controller

Given the incoming feature map $x \in \mathbb{R}^{c \times h \times w}$, the input controller, denoted as $C_I(\cdot)$, performs three operations in the t-th training mini-batch. Firstly, it concatenates x and the old internal memory M_{t-1} from the previous $(t-1)$-th training iteration, and further extracts the internal feature representation via ω_i. Finally, it passes the internal feature representation to the memory controller. The whole process can be formulated as follows:

$$C_I(x) = \phi(x \oplus M_{t-1}; \mathbf{w}_i), \tag{1}$$

where $\oplus$ denotes the concatenate operation. The function $\phi(\cdot)$ denotes the operations (i.e., passing through 3×3 convolutional layer, BN, ReLU, 1×1 convolutional layer, BN and ReLU) illustrated in Figure 3, where $\mathbf{w}_i$ is the corresponding parameter set.

3.2 Memory Controller

Motivated by the recent memory network [Santoro *et al.*, 2016], our designed memory controller enables to flexibly write / read complex and abstract information into the internal memory. Specifically, given the incoming feature map x and the output of $C_I(x)$, the memory controller $C_M(\cdot)$ transforms the old memory M_{t-1} into M_t to make it more representative and general for some intended future use in two steps. Firstly, it generate the memory representation $C_M(x)$ for x by using the internal memory M_{t-1} (updated in the previous mini-batch) with the similarity between x and M_{t-1} as weights. Formally, we have:

$$C_M(x) = \phi\big(x \oplus (\text{clip}_{[\frac{1}{r},r]}(\frac{M_{t-1}}{x}) * x); \mathbf{w}_m\big), \tag{2}$$

where "*" implies the Hadamard product and $\mathbf{w}_m$ denotes the corresponding parameter set of the memory controller. The operation $\text{clip}_{[1/r,r]}(M_{t-1}/x)*x$, inspired from [Ioffe, 2017], denotes only fetching the memory that is similar to x under the empirically constraint of r, i.e., too big or too small value inside M_{t-1}/x will be set as $\frac{1}{r}$ and r, respectively. Intuitively,

this operation is feasible since not all the memory items are beneficial for the current input feature map x. Those items from the internal memory that have small distances to x may be informative to enhance x. Note that, we set r=1 in the beginning to disable the contribution of M_{t-1} since the internal memory may not contain much valuable information. In the middle of the training (e.g., 10000 training iteration), we change r into 3 in all experiments to extract some relevant information from M_{t-1}. Once obtained the memory representation $C_M(x)$ and $C_I(x)$ for the input x, we calculate the new internal memory M_t as:

$$M_t = C_M(x) + C_I(x). \tag{3}$$

3.3 Output Controller

Given the new internal memory M_t, the output controller, denoted as $C_O(\cdot)$, outputs the new memory representation for the input feature map as:

$$C_O(x) = \phi(x \oplus M_t; \mathbf{w}_o), \tag{4}$$

where $\mathbf{w}_o$ is the parameter set for the other operations inside the output controller. Finally, the memory representation C_O is concatenated with the input feature map x and passed to the next neural layer, while the updated internal memory M_t will be further optimized after performing backward propagation for the next training iteration.

3.4 Training and Testing Details

Since all above-mentioned formulations are differentiable, we can directly employ the standard back propagation algorithm [LeCun *et al.*, 1990] to fine-tune the CMB parameters in the training phase. Note that, the internal memory map is randomly initialized at the start of the training and then is uninterruptedly updated without resetting to zero.

In the testing phase, we fix all the parameters inside our CMB, which has been well optimized during the training. Note that, since we expect our CMB to learn to store the informative / representative patterns among samples, the data for both training and testing are randomly shuffled to encourage the internal memory to cache task-specific features via a fully data-driven manner.

3.5 Comparing CMB to ConvLSTM / ConvGRU

Our CMB is entirely different from the ConvLSTM and ConvGRU, although it seems similar to them. The detailed dissimilarities are listed as follows: (i) *Memory Mechanism.* Similar to the LSTM, the ConvLSTM employs the memory cells and hidden states to describe the hidden representation for general-purpose sequence modeling. Therefore, the ConvLSTM requires sequential inputs to obtain a reliable hidden representation during the training and testing phase. The ConvGRU also faces the same issue. Whereas, the internal memory of our proposed CMB adapts to no-sequential inputs by directly fusing with the feature representation of the under-processing sample. Thus, our CMB can be used to store the representative spatial correlations among samples for training; (ii) *Output Generation.* Unlike the ConvLSTM and ConvGRU that consider the updated hidden states as the final output, our CMB directly leverages the newly updated internal memory to process the input feature map to generate outputs. Moreover, our CMB naturally concatenates the output representation and the input feature map to form a residual representation as [He *et al.*, 2016]. This ensures that all the information can be passed directly through the CMB, i.e., our CMB can provide additional feature map enhancement without corrupting the input feature map.

Layer Index	**1**	**2**	**3**	**4**	**5**
Layer Name	conv1_1	conv1_2	max1	CMB1	conv2_1
Channel(kernel-stride)	32(3-1)	32(3-1)	32(2-2)	32	64(3-1)
Layer Index	**6**	**7**	**8**	**9**	**10**
Layer Name	max2	CMB2	conv3_1	conv3_2	max3
Channel(kernel-stride)	64(2-2)	64	128(3-1)	128(3-1)	128(2-2)
Layer Index	**11**	**12**	**13**	**14**	**15**
Layer Name	CMB3	conv4_1	conv4_2	neareast upsampling	concatenate conv3_2
Channel(kernel-stride)	128	256(3-1)	256(3-1)	256	384
Layer Index	**16**	**17**	**18**	**19**	**20**
Layer Name	conv5	neareast upsampling	concatenate conv2_1	conv6	conv7
Channel(kernel-stride)	128(3-1)	128	196	64(3-1)	K(1-1)

Table 1: Details of the proposed articulated pose estimation framework. The scalar K is set according to the number of body joints.

4 CNN with Convolutional Memory Blocks

To clarify the effectiveness of the proposed CMB, we have developed a concise yet powerful framework for articulated pose estimation by embedding the CMB into an hourglass-shape network (see Figure 2), which is inspired by designing principles of the architecture in [Newell *et al.*, 2016]. Regarding the articulated pose estimation as a problem of deep regression as [Tompson *et al.*, 2014], our developed framework takes a depth image as input, and outputs a dense heat map for each joint (e.g., the body part of human). Note that, the heat map denotes a per-pixel likelihood of being the joint. The coordinates of maximum value of the predicted heat map will be treated as the center position of the target body joint.

As illustrated in Table 1, our framework is stacked by ten convolutional layers, three max-pooling layers and two concatenate layers to form an hourglass shape as [Newell *et al.*, 2016]. The kernel size of all the convolutional layers are set to 3×3 with a stride of 1, and three max-pooling layer are set to have 2×2 with a stride of 2. The detailed parameters for our framework can be found in Table 1. It is obvious that our framework contains two downsampling-upsampling steps (resulting in three different scales of feature maps, which are denoted in different colors) to capture the various spatial relationships associated with the pose from the depth image. Therefore, We impose three individual CMBs (i.e., no parameters are shared) to enhance these three kinds of feature maps.

5 Experiments

Dataset Description. We have evaluated the estimation performance of our framework on the newly created Kinect2 Human Pose Dataset (K2HPD) [Wang *et al.*, 2016], which includes about 100K clean depth images with various human poses under three challenging scenarios. We have also used the Invariant-Top View Dataset (ITOP) dataset [Haque *et al.*,

Method	PHR	CPM	SH	IEML	Ours
PDJ (0.05)	26.8	30.0	41.0	43.2	**58.8**
PDJ (0.10)	70.3	58.5	73.7	64.1	**89.0**
PDJ (0.15)	84.7	87.8	84.6	88.1	**94.8**
PDJ (0.20)	91.3	93.6	89.0	91.0	**97.1**
Average	68.3	67.5	72.1	71.6	**84.9**

Table 2: Detailed comparison of the estimation accuracy on the K2HPD benchmark using the PDJ metric.

Method	3DCNN	PHR	CPM	SH	Ours
Head	48.6	83.2	69.9	81.6	**92.9**
Neck	50.9	83.3	71.8	87.3	**97.2**
Shoulders	52.6	82.4	71.1	86.2	**95.2**
Elbows	45.5	71.2	65.7	77.9	**90.4**
Hands	42.1	65.7	65.9	73.4	**86.7**
Writs	42.6	63.4	63.3	72.5	**86.0**
Torso	54.6	80.8	70.4	85.5	**96.6**
Hips	55.0	73.9	66.8	81.2	**93.5**
Knees	50.1	82.7	74.8	86.9	**92.0**
Feet	45.4	81.4	72.7	91.3	**94.3**
Upper Body	47.0	74.2	67.8	79.7	**91.3**
Lower Body	51.5	79.3	70.9	86.0	**93.9**
Full Body	48.9	76.3	69.1	82.3	**92.4**

Table 3: Detailed comparison of the estimation accuracy on the K2HPD benchmark using the PCKh@0.5 metric.

2016], which contains large amount of real-world depth images from two different camera viewpoints by 20 actors performing 15 sequences each. Moreover, we have conducted the experiment on the hand-depth image dataset [Xu and Cheng, 2013] named ASTAR, which consists of 870 depth images of captured by a time-of-flight camera with a data-glove. For a fair comparison on these benchmarks, we follow the same training and testing setting as their officially defined.

Compared Methods. For human pose estimation on the ITOP benchmark, we have compared our framework with Random Forest (RF) [Shotton *et al.*, 2013], Random Tree Walk (RTW) [Jung *et al.*, 2015], Iterative Error Feedback (IEF) [Carreira *et al.*, 2016], and Viewpoint Invariant (VI) [Haque *et al.*, 2016]. On the K2HPD benchmark, we can compare our framework with Inference Embedded Multi-task Learning (IEML) [Wang *et al.*, 2016]. In order to justify the advancement of our framework on depth-based human pose estimation, we have also considered the RGB-based human pose estimation approaches. We have made a quantitative comparison with five RGB-based state-of-the-art methods, i.e., 3DCNN [Grinciunaite *et al.*, 2016], Part Heat-map Regression (PHR) [Bulat and Tzimiropoulos, 2016], Convolutional Pose Machines (CPM) [Wei *et al.*, 2016], and Stacked Hourglass (SH) [Newell *et al.*, 2016]. To directly apply the state-of-the-art RGB-based network architectures to handle the depth data (having a single channel), we need to reduce the input channel of these networks from 3 to 1 and train them from scratch on the above-mentioned benchmarks.

Evaluation Metric. To measure the accuracy of predicting human body joints, we employ the popular Percent of Detected Joints (PDJ) metric [Toshev and Szegedy, 2014], Percentage of Correct Key points (PCKh@0.5), and 10cm-rule [Haque *et al.*, 2016] as the evaluation criteria. Specifically, the PDJ metric considers a body joint is correctly estimated only if the distance between the predicted and the true joint position is within a certain fraction of the torso diameter. The PCKh@0.5 metric recognizes the estimated joint position as correct if its distance to the ground truth joint in the image space is within 50% of the head segment length (i.e., the distance between the head and neck joints). The 10cm-rule metric identifies the correct prediction when the distance between the predict joint and the ground truth joint is less than 10cm in the 3D world coordinate defined by Kinect.

Method	RF	RTW	IEF	VI	Ours
Head	63.8	97.8	96.2	**98.1**	97.0
Neck	86.4	95.8	85.2	97.5	**98.5**
Shoulders	83.3	94.1	77.2	**96.5**	75.1
Elbows	73.2	**77.9**	45.4	73.3	64.7
Hands	51.3	70.5	30.9	68.7	**85.0**
Torso	65.0	93.8	84.7	85.6	**94.5**
Hips	50.8	80.3	83.5	72.0	**88.4**
Knees	65.7	68.8	81.8	69.0	**84.2**
Feet	61.3	68.4	80.9	60.8	**82.9**
Upper Body	70.7	**84.8**	61.0	84.0	80.6
Lower Body	59.3	72.5	82.1	67.3	**86.5**
Full Body	65.8	80.5	71.0	77.4	**83.4**

Table 4: Comparison of the estimation accuracy on the ITOP (front-view) using the 10cm-rule metric.

Method	PHR	CPM	SH	IEML	Ours
Time	62	72	56	40	14
Model Size	570	525	418	-	58
Complexity	63.1	85.0	30.7	-	5.4

Table 5: Comparison of the average running time (milliseconds per image), model size (MB) and model complexity (GFLOPs) on the K2HPD benchmark. Note that, '-' denotes the result is not available.

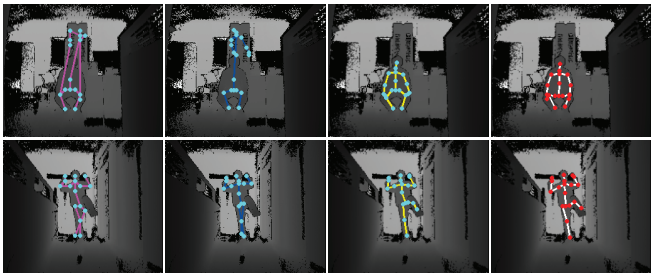

Figure 4: Qualitative comparison between our framework and the compared state-of-the-art methods on the K2HPD benchmark. The estimated joints are directly shown in the images. It is obvious that our framework can obtain much more accurate estimations than the compared methods. Best viewed in color.

Implement Details. All our experiments are carried out on a desktop with Intel 3.4 GHz CPU and NVIDIA GTX-980Ti GPU. In order to reduce overfitting, we employ the image horizontal-flipping and rotating strategy to augment the training data. As for the training process, we train our model from scratch by Adam optimizer [Kendall and Cipolla, 2016] with the batch size of 16 and the initial learning rate of 0.00025, β_1=0.9, β_2=0.999. An exponential learning rate decay is applied with a decay rate of 0.95 every 1000 training iterations.

5.1 Results and Comparisons

Table 2 demonstrates the comparison results under both the PDJ and PCKh@0.5 on the K2HPD benchmark. As one can see from Table 2, our framework significantly outperforms all the compared state-of-the-art methods under every normalized precision threshold. This validates that our framework is more suitable for estimating human poses from depth data, compared with those state-of-the-art methods for RGB data. The similar significant performance gain can also be observed in Table 3. Specifically, our framework has obtained the highest accuracy on all types of body joints. Some visual comparison results are shown in Figure 4.

The comparison results on the ITOP dataset (front-view) are illustrated in Table 4. As shown, it is obvious that our framework dramatically surpasses all the compared methods on the estimation accuracy of full body by a clear margin, and is significantly superior to others. Note that, since the prediction of our framework is 2D, we build upon the heat map layer by two fully connected layers with 1024 neurons to regress 3D predictions.

To further evaluate our framework on the small-scale data, we have conducted the 3D hand pose estimation experiment on the ASTAR benchmark, which only provides 435 images for training. The median/mean joint error is regarded as evaluation metric, and obtained after submitting the estimated hand pose coordinates to the official evaluation server. The median/mean joint errors of our framework is 10.02/11.42mm, and is nearly 100% better than the current state-of-the-art method [Xu *et al.*, 2015], which only achieves 21.1/22.7 mm. This validates that our framework is general to the small-scale hand pose estimation task.

To verify the contribution of our CMB to make network lightweight, we further quantitatively perform real-time analysis of our framework and the compared approaches (except for 3DCNN, which requires sequential data) on the K2HPD dataset. As demonstrated in Table 5, our method runs about 70fps, and performs about 4 times faster than the best of the compared approaches. The reason is that the model size of our method is only 58MB with the complexity 5.4GFLOPS, nearly 10, 6 and 5 times smaller than the compared PHR, CPM and SH, respectively. This demonstrates that the superior performance of our framework, thanks to the employed lightweight architecture with moderate parameters.

5.2 Component Analysis

To perform the detailed component analysis of our framework, we have conducted the following experiment on the K2HPD benchmark to validate the contributions of the introduced CMB. Specifically, we have discarded all the CMBs inside our proposed framework, and directly employ the convolutional layers to estimate human poses. This variant version of our framework reflects the pure performance of the fully convolutional networks, and is denoted as "Ours w/o CMB". We have also conducted two variants of our framework by replacing the convolutional memory block with ConvLSTM [Shi *et al.*, 2015] and ConvGRU [Ballas *et al.*, 2016], and denote them as "Ours w/ ConvLSTM" and "Ours w/ ConvGRU", respectively. Note that, these two variants require sequential data for training and testing, i.e., the action sequence for each subject are sequentially fed into the network without shuffling. Given the same sequential data, we have also trained and tested our framework (denoting as "Ours w/ Sequence". To further analyze the performance of the internal memory capacity of the CMB, we have modified the feature map number of the internal memory inside the CMBs. Specifically, the "Ours w/ Half" denotes the variant of Ours that the feature map number is decreased to half of its original, while the "Ours w/ Double" denotes the channel number is doubled. The original memory capacity of CMB is the same as the channel of the input feature map from the previous layer.

Method	Ours w/o CMB	Ours w/ ConvGRU	Ours w/ ConvLSTM	Ours w/ Sequence	Ours w/ Half	Ours w/ Double	Ours
Head	91.9	94.3	91.2	97.2	97.2	**97.7**	92.9
Neck	95.4	92.7	95.3	97.1	96.9	**97.9**	97.2
Shoulders	92.7	91.1	92.8	95.2	94.9	**96.0**	95.2
Elbows	84.6	85.6	86.3	90.7	90.8	**90.9**	90.4
Hands	81.5	80.1	82.3	86.7	86.7	**87.1**	86.7
Writs	79.9	80.6	81.8	85.8	85.8	**87.0**	86.0
Torso	89.2	92.3	92.8	96.0	95.6	**97.0**	96.6
Hips	74.3	86.3	85.2	92.3	91.8	**93.8**	93.5
Knees	80.8	78.4	85.2	91.5	90.5	90.2	**92.0**
Feet	90.2	87.0	90.9	93.9	93.7	93.7	**94.3**
Upper Body	87.4	87.0	88.1	91.7	91.6	**92.3**	91.3
Lower Body	82.7	85.6	88.0	93.3	92.7	**93.5**	93.9
Full Body	85.4	86.4	88.1	92.3	92.1	**92.8**	92.4

Table 6: Detailed comparison of the estimation accuracy for component analysis on the K2HPD benchmark using the PCKh@0.5 metric. The entries with the best values for each row are bold-faced.

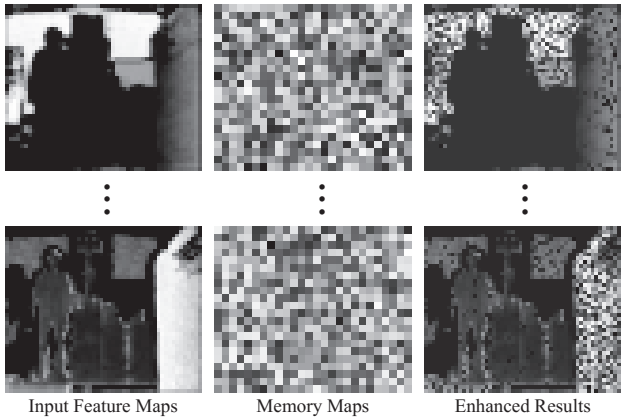

Figure 5: The visualization of the selected input feature maps, its corresponding memories, and the enhanced results by the memory. As shown, for each input feature map, several high response regions belonged to the background are heavily suppressed, while those of human body are mainly preserved.

Table 6 demonstrates the PCKh@0.5 comparison results. As one can see from Table 6, there lies a significant performance gap between our method and all the competing methods. This highlights the superiority of our proposed CMB. Thanks to the proposed CMB, the representation of the convolutional layers inside our framework has been significantly enhanced. Moreover, without requiring sequential training data, our framework even performs about 6% and 4% better than the Ours w/ convGRU and Ours w/ ConvLSTM, respectively. More importantly, our framework achieves sim-

ilar performance as the Ours w/ Sequence. This proves the superior performance of our CMB in storing the representative features for human pose estimation from non-sequential depth data. As one can see from Table 6, the estimation accuracy of our framework slightly increase from 92.1% to 92.8% as the memory capacity inside the CMB grows from Ours w/ Half to Ours w/ Double. The reason may be that larger memory capacity can store richer implicit structural features for convolutional layer augmentation. However, larger memory capacity can also bring much more computational cost (nearly double). Therefore, we employ the current capacity to achieve a trade-off between the accuracy and efficiency.

Moreover, since the memory map contributes to store the abstraction patterns for reusing the rich implicit convolutional structures of CNNs and spatial correlations among the training samples, we demonstrate that how the internal memory enhances the input feature maps in Figure 5. It is obvious that the memory maps can enhance the input feature maps by suppressing their high response regions of the background and preserving their human body regions. The reason is that the background regions can not be well represented by the internal memory. This demonstrates the effectiveness of our proposed CMB.

5.3 Evaluation with Insufficient Training Data

To further explore the effectiveness of our framework under the insufficient training data setting, we have fine-tuned our framework with different percentages of training data on the K2HPD benchmark using the PCK@0.5 metric. We have also compared our framework with Ours w/o CMB to validate the contribution of the proposed CMB. For a fair comparison, we have evaluated a variant of our method (denoted as "Ours w/ Convs") by replacing CMB with convolutional layers in a fair number of parameters and similar architecture. Specifically, we directly duplicate the input feature map to replace the internal memory for each convolutional controller.

The detailed estimation accuracy with standard deviation of 5 repeated trails using the PCK@0.5 metric is listed in Figure 6. As the percentage of training data increases, the increased estimation accuracy can be gradually obtained. However, the accuracy of our framework obtains a steady growth with small deviation from 10% to 100% of training data, while that of Ours w/o CMB increases sharply with large deviation. Especially, under the 10% training data setting, our framework performs nearly 11% better than Ours w/o CMB. Moreover, Ours w/ Convs performs significantly better than Ours w/o CMB due to the additional introduced convolutional parameters. However, our framework still consistently outperforms Ours w/ Convs by clear margins when the percentage of used training data is small. This comprehensively validates the significant contribution of the CMB on improving learning efficiency when given insufficient training data.

6 Conclusion

This paper presented a novel memory network module called Convolutional Memory Block (CMB) for improving the learning efficiency and representation of CNNs on still depth images, which lack of color / texture information. The CMB

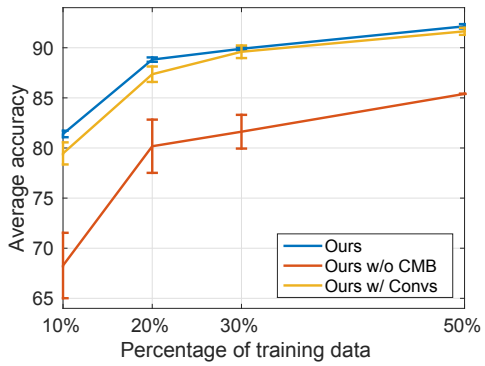

Figure 6: Experimental study on the average estimation accuracy with standard deviation under various percentages of training data from K2HPD using PCK@0.5 metric.

is designed to enable the memory mechanism of convolutional layers by leveraging an internal memory with three specific controllers to store the rich representative structural features and spatial correlations among training samples. Based on the proposed CMB, we have developed a concise yet powerful articulated pose estimation framework, which employs three CMBs to enhance its three different scales of convolutional feature maps. Extensive experiments validated the effectiveness and efficiency of our CMB and framework. In the future, we will extend our CMB to support depth video data for other tasks, e.g., human action/activity recognition.

Acknowledgments

This work was supported in part by the Hong Kong Polytechnic University's Joint Supervision Scheme with the Chinese Mainland, Taiwan and Macao Universities (Grant no. G-SB20). This work was also supported in part by National Natural Science Foundation of China (NSFC) under Grant U1611461 and Grant 61702565, in part by Science and Technology Planning Project of Guangdong Province of No.2017B010116001, in part by Guangdong "Climbing Program" Special Funds under Grant pdjhb0010, in part by NSFC-Shenzhen Robotics Projects under Grant U1613211, and the Fundamental Research Funds for the Central Universities. We thanks a lot for the pleasant discussion with Guangrun Wang.

References

[Ballas *et al.*, 2016] N. Ballas, L. Yao, C. Pal, and A. Courville. Delving deeper into convolutional networks for learning video representations. In *ICLR*, 2016.

[Bulat and Tzimiropoulos, 2016] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, pages 717–732, 2016.

[Carreira *et al.*, 2016] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feed- back. In *CVPR*, page 4733–4742, 2016.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares,

Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[Graves *et al.*, 2014] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. In *arXiv:1410.5401*, 2014.

[Graves *et al.*, 2016] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwinska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, and et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538:471–476, 2016.

[Grinciunaite *et al.*, 2016] Agne Grinciunaite, Amogh Gudi, Emrah Tasli, and Marten den Uyl. Human pose estimation in space and time using 3d cnn. In Gang Hua and Hervé Jégou, editors, *ECCV Workshops*, 2016.

[Haque *et al.*, 2016] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Li Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In *ECCV*, pages 160–177, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016.

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

[Ioffe, 2017] Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1945–1953, 2017.

[Jung *et al.*, 2015] Ho Yub Jung, Soochahn Lee, Yong Seok Heo, and Il Dong Yun. Random tree walk toward instantaneous 3d human pose estimation. In *CVPR*, pages 2467–2474, 2015.

[Kendall and Cipolla, 2016] A. Kendall and R. Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *ICRA*, pages 4762–4769, 2016.

[LeCun *et al.*, 1990] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, and D. Henderson. Handwritten digit recognition with a backpropagation network. In *NIPS*, pages 396–404, 1990.

[Liang *et al.*, 2016] Xiaodan Liang, Xiaohui Shen, Donglai Xiang, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with local-global long short-term memory. In *CVPR*, pages 3185–3193, 2016.

[McColl *et al.*, 2011] Derek McColl, Zhe Zhang, and Goldie Nejat. Human body pose interpretation and classification for social human-robot interaction. *IJSR*, 3(3):313, Jun 2011.

[Na *et al.*, 2017] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *ICCV*, pages 677–685, 2017.

[Newell *et al.*, 2016] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016.

[Park *et al.*, 2017] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *CVPR*, pages 6432–6440, 2017.

[Santoro *et al.*, 2016] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, pages 1842–1850, 2016.

[Shi *et al.*, 2015] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, pages 802–810, 2015.

[Shotton *et al.*, 2011] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, pages 1297–1304, 2011.

[Shotton *et al.*, 2013] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Commun. ACM*, 56(1):116–124, 2013.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[Tompson *et al.*, 2014] Jonathan Tompson, Arjun Jain, Yann Lecun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, pages 1799–1807, 2014.

[Toshev and Szegedy, 2014] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014.

[Wang *et al.*, 2016] Keze Wang, Shengfu Zhai, Hui Cheng, Xiaodan Liang, and Liang Lin. Human pose estimation from depth images via inference embedded multi-task learning. In *ACM MM*, pages 1227–1236, 2016.

[Wei *et al.*, 2016] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016.

[Weston *et al.*, 2015] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *ICLR*, 2015.

[Xu and Cheng, 2013] C. Xu and L. Cheng. Efficient hand pose estimation from a single depth image. In *ICCV*, pages 3456–3462, 2013.

[Xu *et al.*, 2015] Chi Xu, Ashwin Nanjappa, Xiaowei Zhang, and Li Cheng. Estimate hand poses efficiently from single depth images. *IJCV*, 116(1):21–45, 2015.

[Zhang *et al.*, 2018] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 2018.