Progressive Bi-C3D Pose Grammar for Human Pose Estimation

Lu Zhou,[1,2] Yingying Chen,[1,2] Jinqiao Wang,[1,2] Hanqing Lu[1,2]
[1]National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{lu.zhou, yingying.chen, jqwang, luhq}@nlpr.ia.ac.cn

Abstract

In this paper, we propose a progressive pose grammar network learned with Bi-C3D (Bidirectional Convolutional 3D) for human pose estimation. Exploiting the dependencies among the human body parts proves effective in solving the problems such as complex articulation, occlusion and so on. Therefore, we propose two articulated grammars learned with Bi-C3D to build the relationships of the human joints and exploit the contextual information of human body structure. Firstly, a local multi-scale Bi-C3D kinematics grammar is proposed to promote the message passing process among the locally related joints. The multi-scale kinematics grammar excavates different levels human context learned by the network. Moreover, a global sequential grammar is put forward to capture the long-range dependencies among the human body joints. The whole procedure can be regarded as a local-global progressive refinement process. Without bells and whistles, our method achieves competitive performance on both MPII and LSP benchmarks compared with previous methods, which confirms the feasibility and effectiveness of C3D in information interactions.

1 Introduction

Human pose estimation serves as one of the fundamental research directions in computer vision, which aims at locating the joints (head, shoulders, elbows, writsts, knees, ankles, etc.) positions given images that contain various human poses. Human pose estimation has become a significant basis for many other vision tasks. However, human pose estimation still has difficulty in accurate location due to the occlusion, complex human pose gesture, scale variation, and foreshortening.

Human pose estimation has achieved significant progress due to the development of the deep convolutional networks. (Wei et al. 2016) refined the estimated pose stage by stage by enlarging the receptive field of the model. The repeated conv-deconv process proposed in (Newell, Yang, and Deng 2016) captured various scale features and boosted the performance by a large margin. Although considerable enhancements have been made by the techniques proposed, it still remains difficult for the network to obtain more precise predictions. Previous works seldom take consideration of prior knowledge, which results in pool performance under the circumstances of the severe occlusion and complex gesture as shown in Figure 1.

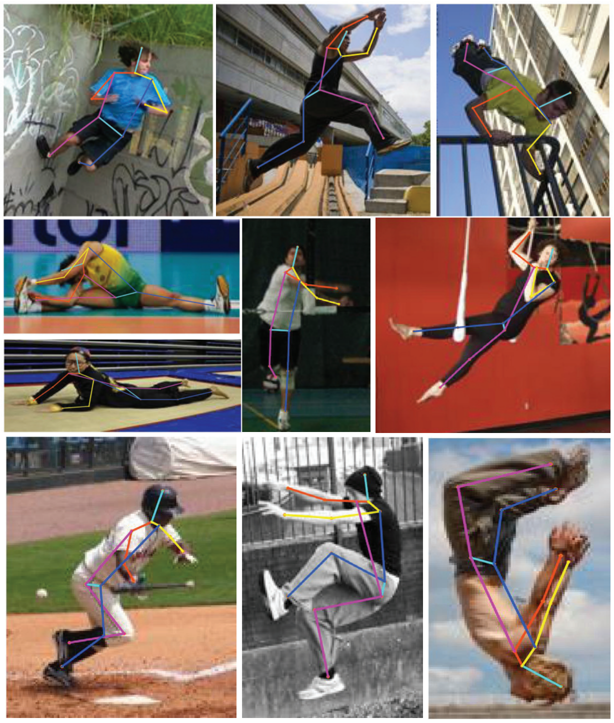

Figure 1: Visualization of the estimated results on the LSP dataset. Our method achieves promising estimation results on the images which cover various pose changes.

High-level semantic information flow among human joints, which can also be regarded as a message passing process, enriches the context of individual features. The wrongly predicted joints can be easily rectified under the assistance of correct predictions and problems mentioned above can be settled thereof. To achieve this goal, we take advantage of the pose grammar mechanism to propagate the message among the feature maps of corresponding human joints.

In this paper, we advance a progressive grammar network to broadcast information among human joints. Firstly, a multi-scale Bi-C3D kinematics pose grammar is imposed to promote local information interactions. We then harness the global sequential Bi-C3D pose grammar to further encourage long-range information flow.

We develop a Bi-C3D grammar module which captures the kinematics relationships among human joints. For instance, we can build grammar submodule which is related to the wrist, elbow, and shoulder to promote the message passing among them. Instead of adopting the Bi-LSTM which causes large memory consumption, we take use of Bi-C3D which shows great advantages in spatiotemporal feature learning to promote the message passing. Our experiments demonstrate that Bi-C3D based grammar learning brings in non-trivial improvements compared with the baseline and reduces considerable memory consumptions during both training and inference. Besides, we experiment with the dilated operation in that enlarging convolution kernel size to cover the spatial range of neighboring joints consumes much more parameters whereas the increase is not that obvious.

Intrinsically, features with different resolutions enjoy different levels semantic information. Low resolution features capture global information which depicts the whole skeleton knowledge of the human body. However, the high resolution features learn more spatial details which pose great significance in accurate localization. In this paper, we extend the grammar network to a multi-scale framework to promote multi-level message passing. Different parameter settings are adopted across different scales to ensure enough spatial coverage between neighboring joints. Multi-scale grouped supervision is also imposed here. Furthermore, we fuse the features after message passing to exploit multi-scale contextual information. The fused features enhance the robustness of the network when faced with the scale changes of the whole input human and corresponding human joints.

To exploit long-range dependencies among the human joints, we propose a global sequential grammar learned with Bi-C3D to capture the semantic information of the whole body not only the semantically related joints. Firstly, we rearrange the order of human joints according to spatial relevance instead of tree or loopy structure. The holistic structure information can be learned from Bi-C3D module which shows great advantages in promoting long-range message interactions.

We adopt hourglass model as our basic structure due to its outstanding performance in human pose estimation, human parsing and object detection.

The contributions of this work are summarized as follows:

- We propose a multi-scale Bi-C3D kinematics grammar learning framework to enforce information flow of different granularity levels. The kinematics grammar mainly captures the associations and adjacency among locally related body joints. To the best knowledge of ours, this is the first attempt to apply C3D module to learn the domain-specific knowledge of the human pose estimation. Additionally, fusion of multi-scale context-enriched feature maps enhances the scale invariance of the network.
- To capture long-range dependencies of the human body joints, we advocate a novel sequential grammar module to propagate long-range message passing instead of the tree or the loopy structure. It is convenient to implement with Bi-C3D and proves effective.
- Combining the multi-scale Bi-C3D kinematics grammar and global sequential grammar together executes progressive information exchange among the human joints. Our experiments confirms the reasonability and validity of progressive integration of the two grammars.

2 Related Work

2.1 Human Pose Estimation

Human pose estimation has achieved great progress due to the development of the DCNNs. DeepPose(Toshev and Szegedy 2014) tried to estimate the body joints by coordinates regression and it was one of the first attempts to use the deep convolutional features for inference. However, coordinate regression based methods are difficult to converge. (Tompson et al. 2014; 2015) attempted to take advantage of MRF to eliminate the false positive predictions. (Chu et al. 2016) proposed a CRF-CNN framework to model the inherent structure of the human body in a probabilistic way. (Lifshitz, Fetaya, and Ullman 2016) utilized the voting scheme to improve the whole performance. (Wei et al. 2016) modified the traditional pose machines into a deep convolutional framework and the repeated intermediate supervision mechanism promoted gradient flow and boosted the performance. (Newell, Yang, and Deng 2016) adopted the conv-deconv, encoder-decoder mode to capture different scales information and the multiple intermediate supervision mechanism was also enforced in the framework. Subsequent works such as (Chou, Chien, and Chen 2017; Chu et al. 2017; Yang et al. 2017; Chen et al. 2017; Peng et al. 2018; Ning, Zhang, and He 2018) all employed hourglass as their backbone and explored more effective framework. In this paper, we propose a grammar network based on C3D to promote the message passing instead of probabilistic graph.

2.2 Grammar Model

Grammar based model has shown its effectiveness in various domains. (Han and Zhu 2009) proposed a stochastic grammar model to solve the image parsing problem. (Xiaohan Nie, Wei, and Zhu 2017) proposed to build the kinematic grammar with skeleton-LSTM and patch-LSTM. A hierarchical attributed grammar network was proposed in (Fang et al. 2018) to enforce high-order constraints over 3D human poses. (Wang et al. 2018) leveraged two fashion grammars to encode the high-level human knowledge for the fashion landmark detection. Different from the methods above, we propose a local-global multi-scale grammar learning network with C3D instead of largely memory-consumed LSTM.

3 Method

The whole framework is illustrated in Figure 2. In this section, we briefly introduce our progressive Bi-C3D gram-

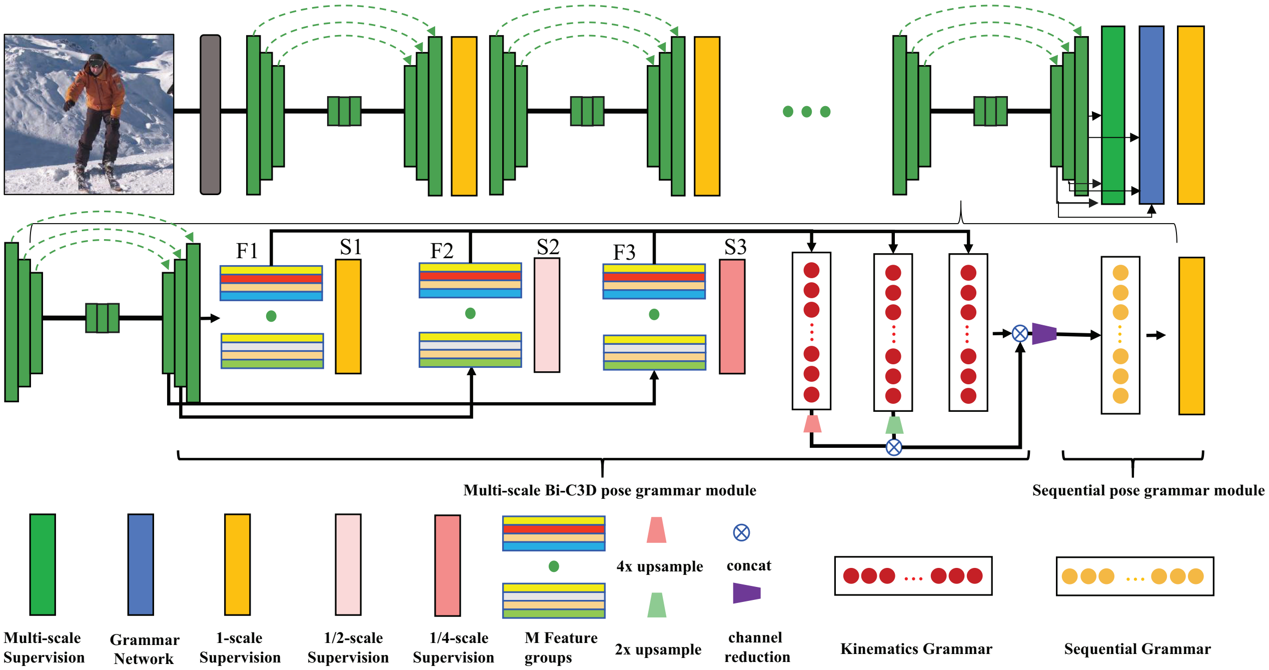

Figure 2: The proposed multi-scale Bi-C3D pose grammar network. We accept 8-stack hourglass network as backbone and the grammar module which encodes human body dependencies and relations is embedded into the last stack of the hourglass model. The multi-scale Bi-C3D pose grammar network consists of the multi-scale Bi-C3D pose kinematics grammar and global sequential grammar.

mar network. The whole procedure can be regraded as a local-global progressive refinement process. The multi-scale Bi-C3D kinematics grammar module mainly captures local constraints amid locally correlated joints. The long-term sequential pose grammar concentrates on developing global configuration of the human pose. Combining the grammars mentioned above realizes step-by-step refinement and proves effective on the human key point localization task.

3.1 Basic Structure

In order to build a solid foundation for grammar learning, we employ hourglass model which repeats the down-sampling and up-sampling operations to extract high-level 2D pose features. The grammar module is embedded into the backbone to capture the high-order context. However, there exist some differences in details. Take 1-scale features for example, in practice, instead of feeding the features $F1$ into the 1×1 convolution for final prediction, the features are divided into M branches for inference, where M represents the number of joints. Each branch of the features is convolved with 1×1 convolution kernel to output heatmaps of the corresponding human joints. We take advantage of these M groups features to build the kinematics grammar.

3.2 Bi-C3D Kinematics Grammar

The proposed Bi-C3D kinematics grammar depicts constraints of the kinematically connected human joints. In this section, we define 5 kinematics grammars where local refinement is carried on:

$$\begin{aligned}
G_1^K &: r.ankle \longleftrightarrow r.knee \longleftrightarrow r.hip,\\
G_2^K &: l.ankle \longleftrightarrow l.knee \longleftrightarrow l.hip,\\
G_3^K &: r.wrist \longleftrightarrow r.elbow \longleftrightarrow r.shoulder,\\
G_4^K &: l.wrist \longleftrightarrow l.elbow \longleftrightarrow l.shoulder,\\
G_5^K &: head \longleftrightarrow neck \longleftrightarrow thorax \longleftrightarrow pelvis.
\end{aligned} \tag{1}$$

The grammar described above establishes the relationships of human joints in kinematic chain, which reflects human anthropomorphic and anatomical constraints. Geometric construction can be found in Figure 3.

Message passing process of the grammar module is implemented by bi-directional (forward/backward) C3D (Bi-C3D), which naturally supports chain-like structures. Take G_1^K for example, suppose the feature tensors of right ankle, right knee, right hip as follows:

$$\begin{aligned}
X_{rank} &\in R^{N \times C_{in} \times H \times W},\\
X_{rknee} &\in R^{N \times C_{in} \times H \times W},\\
X_{rhip} &\in R^{N \times C_{in} \times H \times W},
\end{aligned} \tag{2}$$

where N represents the batch size, C, H, W represent the channel number, height, width of the corresponding feature tensor. For the forward pass, we concatenate $X_{rank}, X_{rknee}, X_{rhip}$ along the depth dimension sequentially and for the backward pass, we concatenate

$X_{rank}, X_{rknee}, X_{rhip}$ along the depth dimension in a reverse order, that is,

$$G^K_{1\_for} \in R^{N\times3\times C_{in}\times H\times W}, \\ G^K_{1\_back} \in R^{N\times3\times C_{in}\times H\times W}. \tag{3}$$

The resulted grammar features $G^K_{1\_for}, G^K_{1\_back}$ are feeded into the Bi-C3D module to build the chain-like grammar. The Bi-C3D module consists of two groups 3D convolutions. We employ the first group for the forward pass and the second group for the backward pass. Each group is composed of three successive 3D convolution kernels and concrete kernel size can be found in Sec. 3.3. The output features after the message passing can be denoted as

$$O^K_{1\_for} \in R^{N\times3\times C_{out}\times H\times W}, \\ O^K_{1\_back} \in R^{N\times3\times C_{out}\times H\times W}. \tag{4}$$

We acquire the bi-directional context-enriched features by splitting the output features $O^K_{1\_for}, O^K_{1\_back}$ along the depth dimension. Take right ankle for example,

$$X_{rank\_for} \in R^{N\times C_{out}\times H\times W}, \\ X_{rank\_back} \in R^{N\times C_{out}\times H\times W}. \tag{5}$$

We sum the forward context-enriched features of right ankle and backward ones to obtain the final features of right ankle where

$$X_{rank} = X_{rank\_for} + X_{rank\_back}. \tag{6}$$

The grammar of other types can be obtained following the same procedure. The features after the message passing among corresponding human joints incorporate information of the neighboring joints and range of the influence lies on the kernel size of the depth dimension. We set kernel size of depth dimension of the local kinematics grammar as 3.

3.3 Multi-Scale Bi-C3D Kinematics Grammar Network

Hourglass model concentrates on learning multi-scale features via repeated down-up sampling operations. Different resolution features represent contextual information of different scales. To promote multi-scale message exchange and exploit multi-scale contextual information, we embed the kinematics grammar into each scale of the hourglass model as illustrated in Figure 2.

Following the grammar building procedure described in Sec 3.1, Sec 3.2, we firstly enforce multi-scale intermediate supervision explicitly on a single stack of hourglass model. The multi-scale intermediate supervision strengthens the robustness of the network when faced with scale changes and promotes the gradient flow. Multi-scale supervision $S1, S2, S3$ is enforced at the end of each deconvolution layer as shown in Figure 2. Each of the features $F1, F2, F3$ extracted from the end of each de-convolution layer is separated into M branches and each branch is convolved with 1×1 kernel to obtain heatmaps of all the joints across different scales.

Table 1: Different dilated rate settings across multiple scales for different iterations.

	Iter1	Iter2	Iter3
Scale1	2	4	4
Scale1/2	1	2	2
Scale1/4	1	1	1

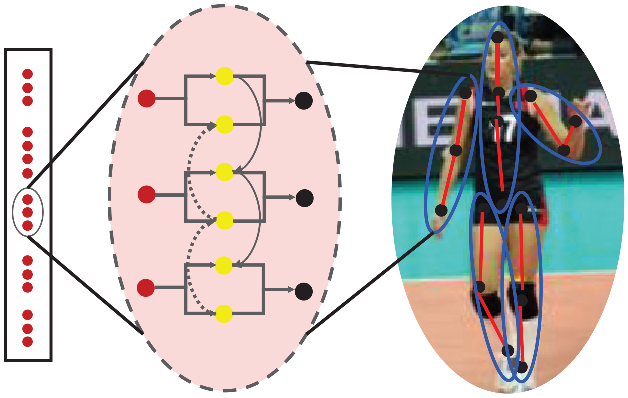

Figure 3: Illustration of the local Bi-C3D kinematics grammar which depicts the knowledge of human body composition. The local Bi-C3D kinematics grammar module is constituted by 5 kinematics sub-grammars as shown by the blue ellipse.

In practice, we adjust the parameter settings across different scales to ensure enough spatial coverage. We perform three successive 3D convolution iterations which exploit 3-by-3-by-3 C3D kernels at scale $1, 1/2, 1/4$ respectively. Instead of applying large convolution kernel to expand the search region of adjacent joints, we take advantage of successive spatially dilated 3D convolutions to perform the message passing. On the one hand, the respective field of neighboring human joints is enlarged to ensure enough information grasping between them. On the other hand, the sparse dilated kernel saves total parameters of the model and reduces the amount of calculation with comparable performance. We set the dilation rate on the spatial (height and width) dimension of the three successive kernels as $\{2, 4, 4\}$ at scale 1, $\{1, 2, 2\}$ at scale $1/2$, $\{1, 1, 1\}$ at scale $1/4$. We set the dilation rate on the depth dimension as 1 across all the situations. Choice of the dilation rate on the spatial dimension can be regarded as a problem of permutations and combinations while it's not the main concern of our work. Concrete configurations can be found in Table 1. We upsample the $1/2$-scale, $1/4$-scale context-enriched features to the same resolution as the 1-scale one and fuse them by concatenation. We reduce the dimension of the fused features from $C_{out} \times 3$ to C_{out} and the details can be found in Figure 2.

3.4 Sequential Grammar

In this section, a novel sequential grammar is proposed to learn long-range dependencies to obtain more accurate pre-

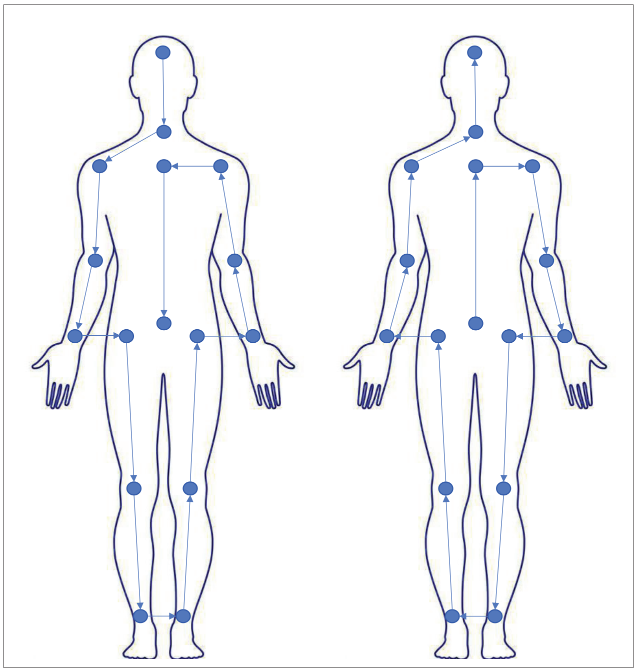

Figure 4: Illustration of the global sequential grammar which further explores higher level context information. There are two directions for the message passing as the arrow shows.

dictions. Multi-scale Bi-C3D sequential grammar network is out of consideration in this work for the efficiency of computation. The features maps of the corresponding joints are concatenated in a predefined

$$\begin{aligned} &head - neck - rshoulder - relbow - rwrist- \\ &rhip - rknee - rankle - lankle - lknee - lhip- \\ &lwrist - lelbow - lshoulder - thorax - pelvis \end{aligned} \tag{7}$$

order to form the inputs of the forward message passing. Concrete illustration of the concatenation order can be found in Figure 4. The procedure of the concatenation follows the same way as in building kinematics grammar. The input of the backward message passing is formed in a reverse order of the forward one:

$$\begin{aligned} R^S_{1\_for} &\in R^{N\times M\times C\times H\times W}, \\ R^S_{1\_back} &\in R^{N\times M\times C\times H\times W}. \end{aligned} \tag{8}$$

We feed $R^S_{1\_for}, R^S_{1\_back}$ into the Bi-C3D module which also consists of two groups 3D convolutions. However, one joint alone is expected to correlate with almost all the other joints. To achieve this goal, more iterations are stacked repeatedly. In practice, we perform 6 successive iterations which also employ 3-by-3-by-3 C3D dilated kernels. We set the dilated rate at spatial dimension, depth dimension as $\{4, 1\}$ respectively across all the 6 successive iterations. We take advantage of long-term sequential modeling and effective message exchange natures of C3D instead of learning the spatiotemporal features in this work.

3.5 Brief Introduction of Other Implementations

In this section, we will give a brief introduction of the implementation of ConvLSTM. To keep the same power of features expressions with C3D implementation, the channel number of each human joint is set as 16 as well. The features involved are propagated bidirectionally and forward information interaction is as follows:

$$\begin{aligned} f_i^f &= \sigma(b_f^f + U_f^f * x_i^f + W_f^f * h_{i-1}^f), \\ i_i^f &= \sigma(b_{in}^f + U_{in}^f * x_i^f + W_{in}^f * h_{i-1}^f), \\ o_i^f &= \sigma(b_o^f + U_o^f * x_i^f + W_o^f * h_{i-1}^f), \\ c_i^f &= tanh(b^f + U^f * x_i^f + W^f * h_{i-1}^f), \\ s_i^f &= f_i^f \circ s_{i-1}^f + i_i^f \circ c_i^f, \\ h_i^f &= o_i^f \circ tanh(s_i^f), \end{aligned} \tag{9}$$

where W, b represent convolution kernels and biases seperately, $*, \circ$ indicate convolution and Hadamard product. In Equation 9, i indicates the sequential order. For comparison, we also adopt dilated convolution and three iterations to cover enough spatial range as done in C3D experiments. The backward information flow can be formulated as in Equation 9 as well. Instead of fusing the bi-directional features inside the inner nodes of LSTM, we incorporate bi-directional information after three iterations message passing. Feature of corresponding human joint is represented as

$$h_i = h_i^f + h_i^b. \tag{10}$$

3.6 Training and Inference

The progressive Bi-C3D pose grammar module is embedded into the last stack hourglass model as shown in Figure 2. The first few hourglass modules where grammar module is not included are only enforced with 1-scale supervision as done in previous work (Newell, Yang, and Deng 2016). We denote ground-truth locations of the human joints as $x = \{x_m\}_{m=1}^M$, where M represents the number of the human joints. We generate the ground-truth score maps $S = \{S_m\}_{m=1}^M$ by enforcing gauss distributions around the ground-truth locations with kernel size 7. The loss of the first few hourglass modules can be denoted as

$$L1 = \sum_{n=1}^{N-1} \sum_{m=1}^{M} ||S_m - \tilde{S}_m||^2, \tag{11}$$

where N represents the stage number of the hourglass model.

Multi-scale supervision is involved in the last stack hourglass module. The multi-scale supervision consists of 3 different scales$(1, 1/2, 1/4)$ supervisions and we finally enforce 1-scale supervision on the context-enriched feature maps after the message passing process to obtain the final prediction. The ground truth score maps of the ith scale supervision can be denoted as $S^i = \{S_m^i\}_{m=1}^M$. Loss for the last stack hourglass model can be formed as

$$L2 = \sum_{i=1}^{J} \sum_{m=1}^{M} ||S_m^i - \tilde{S}_m^i||^2 + \sum_{m=1}^{M} ||S_m - \tilde{S}_m||^2, \tag{12}$$

where J represents the scale number and $\tilde{S}^2_m, \tilde{S}^3_m$ are downsampled by $\tilde{S}^1_m$ with max-pooling operation. Overall loss function can be denoted as

$$L = L1 + L2. \tag{13}$$

During inference, unary maps are taken from the last predictions. The last predictions are obtained from features decorated by the progressive grammar module which are armed with strong semantics. We take the positions with the maximum scores without other techniques as our final prediction $\tilde{x} = \{\tilde{x}_m\}_{m=1}^M$

$$\tilde{x}_m = argmax \tilde{S}_m \tag{14}$$

4 Experiments

4.1 Dataset

The experiments are carried on two widely applied benchmarks MPII (Andriluka et al. 2014) and LSP (Johnson and Everingham 2010). MPII dataset contains about 25k images with 40k annotated samples. There are 16 labeled human joints and 14 of them are expected to be evaluated. We conduct our experiments on the 25925 training images and 2958 valid images. LSP dataset contains the original LSP dataset and the extended version, where 11k training images and 1k test images are involved.

4.2 Experiment Settings

We conduct all our experiments on the platform of Pytorch. Learning rate is set as 5e-4 at the beginning and dropped by 10 at 150 epoch and 170 epoch respectively. We utilize the RMSprop (Tieleman and Hinton 2012) algorithm to update the parameters of the model. We crop the image to the size of 256×256 and person expected to be estimated is located at the center of the cropped patch with roughly the same scale. We rotate the cropped patch by ± 30 and scale the image by a random number. Random color jittering, shearing and flipping are involved as well. Six-scale (0.8,0.9,1.0,1.1,1.2,1.3) image pyramids combined with flipping are adopted during testing. The grammar module is appended at the end of the eight-stack hourglass model.

4.3 Experiments Results

Our method achieves competitive performance on both of the datasets. Our approach achieves 92.5% PCKh score at the threshold of 0.5 on the MPII pose dataset. Concrete details of the final results can be found in Table 2. In contrast, our method surpasses the baseline by 2.2% and 2.8% on the challenging parts ankle and wrist. We incorporate the MPII split into the LSP training set when conduct training process on LSP dataset as done in previous works. Average PCK score at the threshold of 0.2 can be found in Table 3 where person-centric annotation is involved. Our approach achieves 94.8% PCK0.2 score and performs better compared with all the previous methods. According to Table 3 we can find that our method exceeds previous methods on all human parts.

Table 2: Evaluation results using PCKh@0.5 as measurement on the MPII dataset

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
(Tompson et al. 2014)	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
(Tompson et al. 2015)	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
(Hu and Ramanan 2016)	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
(Lifshitz et al. 2016)	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
(Rafi et al. 2016)	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
(Sun et al. 2017b)	97.5	94.3	87.0	81.2	86.5	78.5	75.4	86.4
(Insafutdinov et al. 2016)	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
(Wei et al. 2016)	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
(Bulat et al. 2016)	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
(Newell et al. 2016)	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
(Sun et al. 2017a)	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
(Ning et al. 2016)	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
(Chu et al. 2017)	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
(Liu et al. 2018)	98.4	96.4	92.0	87.9	90.7	88.3	85.3	91.6
(Chou et al. 2017)	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
(Chen et al. 2017)	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
(Yang et al. 2017)	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
(Ke et al. 2018)	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
(Tang, Yu, and Wu 2018)	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
(Sun et al. 2019)	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
ours	98.5	96.9	92.8	**89.3**	91.8	89.5	**86.4**	**92.5**

Table 3: Evaluation results using PCK@0.2 as measurement on the LSP dataset

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
(Rafi et al. 2016)	95.8	86.2	79.3	75.0	86.6	83.8	79.8	83.8
(lifshitz et al. 2016)	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
(Insafutdinov et al. 2016)	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
(Wei et al. 2016)	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
(Bulat et al. 2016)	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
(Sun et al. 2017a)	97.9	93.6	89.0	85.8	92.9	91.2	90.5	91.6
(Chu et al. 2017)	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
(Chen et al. 2017)	98.5	94.0	89.8	87.5	93.9	94.1	93.0	93.1
(Liu et al. 2018)	98.1	94.0	91.0	89.0	93.4	95.2	94.4	93.6
(Yang et al. 2017)	98.3	94.5	92.2	88.9	94.4	95.0	93.7	93.9
(Chou et al. 2017)	98.2	94.9	92.2	89.5	94.2	95.0	94.1	94.0
(Peng et al. 2018)	98.6	95.3	92.8	90.0	94.8	95.3	94.5	94.5
ours	**98.7**	**95.7**	**93.2**	**90.5**	**95.2**	**95.5**	**94.6**	**94.8**

4.4 Ablation Study

We investigate ablation study on the MPII validation set and adopt the two-stack hourglass model as baseline. We incorporate progressive Bi-C3D pose grammar model into the second stack hourglass module.

Effect of the single-scale pose grammar module. From Figure 5, we can see that the single-scale Bi-C3D pose grammar module achieves 88.15% PCKh@0.5 score and boosts the performance by a large margin compared with the pure two-stack hourglass module which only achieved 87.42% PCKh@0.5 score. The C3D pose grammar module builds the relationships among the human pose joints and refines the prediction results.

Effect of the multi-scale pose grammar module. The effect of the multi-scale C3D pose grammar module can be found in Figure 5. Two-scale Bi-C3D pose grammar performs better than single-scale Bi-C3D pose grammar module due to the multi-scale message passing. We adopt the three-scale Bi-C3D pose grammar module at last owing to its superior performance.

Effect of the global sequential grammar. From Figure 5, we can observe that PCKh score of global sequential grammar reaches around 88.16% and surpasses the baseline drastically. Additionally, progressive pose grammar which is constituted by multi-scale pose grammar and sequential grammar surpasses the multi-scale pose grammar. We can

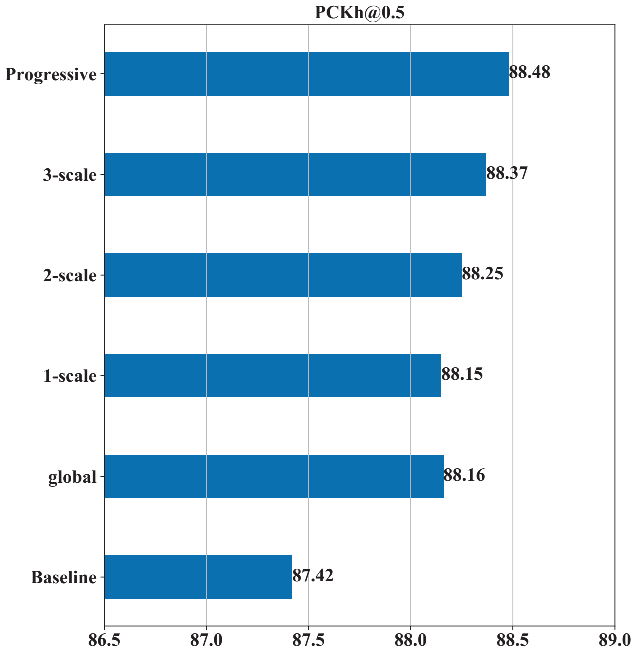

Figure 5: Component investigation on the MPII validation split. The details can be found in Sec. 4.4.

conclude that global sequential grammar can further boost up the performance and information broadcasting among all the human joints proves effective for the final prediction thereof.

Comparison with other techniques. In this section, we compare our single-scale Bi-C3D pose grammar module implemented by D-C3D with other techniques to prove the feasibility and effectiveness of it. From Figure 6, we can find that grammar implemented by D-C3D achieves better results with much less parameters compared with Bi-LSTM pose grammar. During the realistic implementation, the Bi-C3D pose grammar consumes much less memory usage. Additionally, we investigate the effectiveness of D-C3D by replacing D-C3D operation with L-C3D. Though 7 × 7 kernel achieves almost equivalent performance with relatively small field of perception, the parameter number reaches up to 8.82M compared with 6.77M and the speed drops quickly during both the training and inference. The comparison of D-C3D with P-C3D also verifies the effectiveness of spatially dilated C3D operation in information interactions. The P-C3D where dilation is not involved performs poorly owing to insufficient spatial context coverage.

5 Conclusion

This paper has proposed a progressive grammar model constituted by the multi-scale Bi-C3D kinematics grammar module and global sequential grammar module. The paper provides a new perspective to build human pose grammar. First, we develop the multi-scale kinematics C3D pose grammar to capture the multi-scale kinematics information. Kinematics grammar built by Bi-C3D operation takes advantage of the merits of C3D which excels at 3D space message passing. Additionally, a well-designed global sequential grammar which utilizes inherent nature of C3D in long-term relation construction is proposed. The local-global process conducts the message passing progressively and refines the previous predictions. The whole framework can be end-to-end trained and achieves promising results over two public datasets. Our approach verifies the effectiveness and feasibility of C3D in message passing.

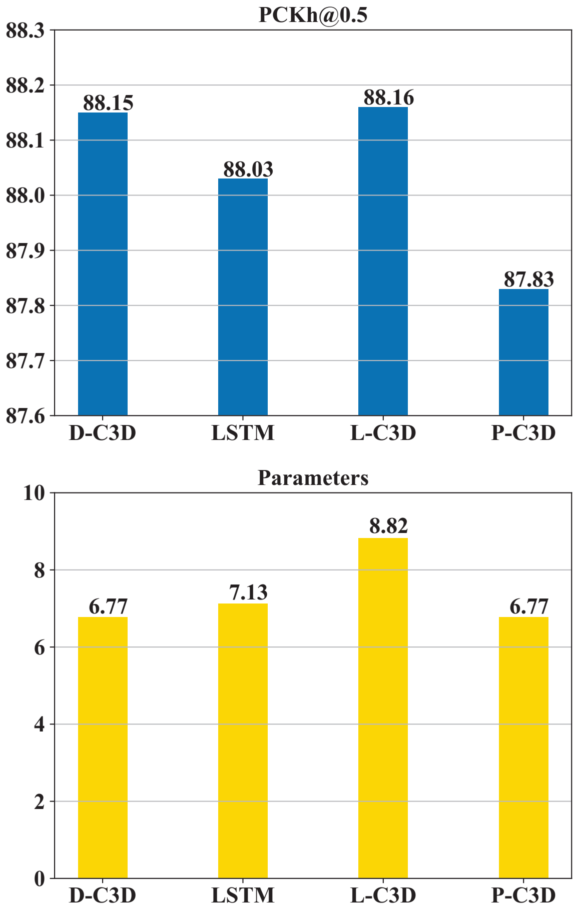

Figure 6: Component investigation on the MPII validation split. D-C3D (Dilated C3D) means spatially dilated C3D operation where the spatial kernel size is set as 3×3, L-C3D (Large kernel C3D) means C3D operation with large spatial kernel size which is set as 7×7, P-C3D (Plain C3D) represents plain C3D operation without dilated settings where the spatial kernel size is set as 3×3. The details can be found in Sec. 4.4.

6 Acknowledgements

This work was supported by National Natural Science Foundation of China 61772527, 61806200, 61976210 and 61702510.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 3686–3693.

Chen, Y.; Shen, C.; Wei, X.-S.; Liu, L.; and Yang, J. 2017. Adversarial posenet: A structure-aware convolutional network for human pose estimation. *CoRR, abs/1705.00389* 2.

Chou, C.-J.; Chien, J.-T.; and Chen, H.-T. 2017. Self adversarial training for human pose estimation. *arXiv preprint arXiv:1707.02439*.

Chu, X.; Ouyang, W.; Wang, X.; et al. 2016. Crf-cnn: Modeling structured information in human pose estimation. In *Advances in Neural Information Processing Systems*, 316–324.

Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A. L.; and Wang, X. 2017. Multi-context attention for human pose estimation. *arXiv preprint arXiv:1702.07432* 1(2).

Fang, H.-S.; Xu, Y.; Wang, W.; Liu, X.; and Zhu, S.-C. 2018. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Proc. of the AAAI Conference on Artificial Intelligence*.

Han, F., and Zhu, S.-C. 2009. Bottom-up/top-down image parsing with attribute grammar. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(1):59–73.

Hu, P., and Ramanan, D. 2016. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5600–5609.

Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; and Schiele, B. 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, 34–50. Springer.

Johnson, S., and Everingham, M. 2010. Clustered pose and nonlinear appearance models for human pose estimation.

Ke, L.; Chang, M.-C.; Qi, H.; and Lyu, S. 2018. Multi-scale structure-aware network for human pose estimation. *arXiv preprint arXiv:1803.09894*.

Lifshitz, I.; Fetaya, E.; and Ullman, S. 2016. Human pose estimation using deep consensus voting. In *European Conference on Computer Vision*, 246–260. Springer.

Liu, W.; Chen, J.; Li, C.; Qian, C.; Chu, X.; and Hu, X. 2018. A cascaded inception of inception network with attention modulated feature fusion for human pose estimation. In *AAAI*.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 483–499. Springer.

Ning, G.; Zhang, Z.; and He, Z. 2018. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia* 20(5):1246–1259.

Peng, X.; Tang, Z.; Yang, F.; Feris, R. S.; and Metaxas, D. 2018. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2226–2234.

Rafi, U.; Leibe, B.; Gall, J.; and Kostrikov, I. 2016. An efficient convolutional network for human pose estimation. In *BMVC*, volume 1, 2.

Sun, K.; Lan, C.; Xing, J.; Zeng, W.; Liu, D.; and Wang, J. 2017a. Human pose estimation using global and local normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 5599–5607.

Sun, X.; Shang, J.; Liang, S.; and Wei, Y. 2017b. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, 2602–2611.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*.

Tang, W.; Yu, P.; and Wu, Y. 2018. Deeply learned compositional models for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 190–206.

Tieleman, T., and Hinton, G. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4(2):26–31.

Tompson, J. J.; Jain, A.; LeCun, Y.; and Bregler, C. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, 1799–1807.

Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; and Bregler, C. 2015. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 648–656.

Toshev, A., and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1653–1660.

Wang, W.; Xu, Y.; Shen, J.; and Zhu, S.-C. 2018. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4271–4280.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4724–4732.

Xiaohan Nie, B.; Wei, P.; and Zhu, S.-C. 2017. Monocular 3d human pose estimation by predicting depth on joints. In *Proceedings of the IEEE International Conference on Computer Vision*, 3447–3455.

Yang, W.; Li, S.; Ouyang, W.; Li, H.; and Wang, X. 2017. Learning feature pyramids for human pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 1290–1299. IEEE.