Deeply Learned Compositional Models for Human Pose Estimation

Wei Tang, Pei Yu and Ying Wu

Northwestern University
2145 Sheridan Road, Evanston, IL 60208
{wtt450, pyi980, yingwu}@eecs.northwestern.edu

Abstract. Compositional models represent patterns with hierarchies of meaningful parts and subparts. Their ability to characterize high-order relationships among body parts helps resolve low-level ambiguities in human pose estimation (HPE). However, prior compositional models make unrealistic assumptions on subpart-part relationships, making them incapable to characterize complex compositional patterns. Moreover, state spaces of their higher-level parts can be exponentially large, complicating both inference and learning. To address these issues, this paper introduces a novel framework, termed as Deeply Learned Compositional Model (DLCM), for HPE. It exploits deep neural networks to learn the compositionality of human bodies. This results in a novel network with a hierarchical compositional architecture and bottom-up/top-down inference stages. In addition, we propose a novel bone-based part representation. It not only compactly encodes orientations, scales and shapes of parts, but also avoids their potentially large state spaces. With significantly lower complexities, our approach outperforms state-of-the-art methods on three benchmark datasets.

1 Introduction

Human pose estimation (HPE) means to locate body parts from input images. It serves as a fundamental tool for several practical applications such as action recognition, human-computer interaction and video surveillance [1]. The most recent HPE systems have adopted convolutional neural networks (CNNs) [2–4] as their backbones and yielded drastic improvements on standard benchmarks [5–9]. However, they are still prone to fail when there exist ambiguities caused by overlapping parts, nearby persons and clutter backgrounds, *e.g.*, Fig. 1.

One promising way to tackle these difficulties is to exploit the compositionality [10, 11] of human bodies. It means to represent a whole body as a hierarchy of parts and subparts, which satisfy some articulation constraints. This kind of hierarchical structure enables us to capture high-order relationships among parts and characterize an exponential number of plausible poses [12]. Based on this principle, compositional models[1] [13, 14] infer poses via two stages, as illustrated in Fig. 2(a). In the bottom-up stage, states of higher-level parts are

[1] We focus on multilevel compositional models in this paper.

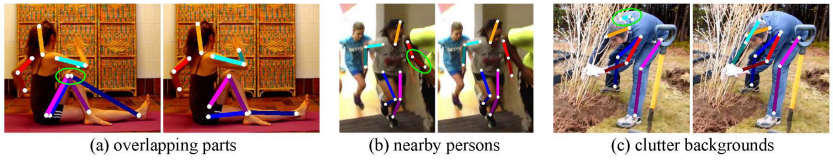

Fig. 1. Pairs of pose predictions obtained by an eight-stack hourglass network [5] (left) and our approach (right). Some wrong part localizations are highlighted by green ellipses. By exploiting compositionality of human bodies, our approach is able to reduce low-level ambiguities in pose estimations. See Fig. 8 for more examples

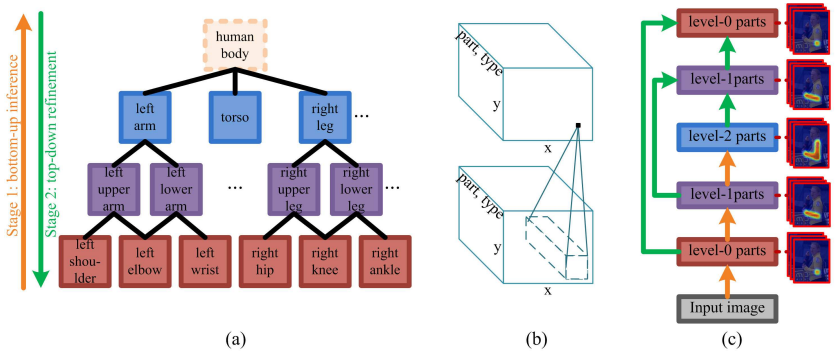

Fig. 2. (a) A typical compositional model of a human body. The pose is estimated via two stages: bottom-up inference followed by top-down refinement. (b) Each tensor represents score maps of several parts. An SLIS function aggregates information from input score maps on a spatially local support to predict output score maps. (c) Overview of our deeply learned compositional model. The orange and green arrows respectively denote SLIS functions modeled by CNNs in bottom-up and top-down stages. The colored rectangles on the left side denote predicted score maps of parts at different semantic levels while the heat maps on the right side represent their corresponding ground truth in the training phase

recursively predicted from states of their child parts. In the top-down stage, states of lower-level parts are refined by their parents' states updated one step earlier. Such global adjustments enable pose estimations to optimally meet the relational constraints and thus reduce low-level image ambiguities. In the last decade, compositional models have been adopted in several HPE systems [12, 15–19] and shown superior performances over their flat counterparts.

However, there are problems with existing compositional models designed for HPE [12, 15–19]. First, they often assume a Gaussian distribution on the subpart-part displacement with the subpart's anchor position being its mean. While simplifying both their inference and learning [20], this assumption generally does not hold in real scenarios, *e.g.*, distributions of joints visualized in [21–23]. Thus, we argue it is incapable to characterize the complex compositional relationships among body parts. Second, a set of discrete *type* variables are often used to

model the compatibility among parts. They not only include the orientation and scale of a part but also span semantic classes (a straight versus bended arm). As the distinct types of a part can be as many as the different combinations of all its children's types, state spaces for higher-level parts can be exponentially large. This makes both computation and storage demanding. Third, when the compositional structure has loops, approximate inference algorithms must be used. As a result, both the learning and testing will be adversely affected.

To address these issues, this paper introduces a novel framework, termed as Deeply Learned Compositional Model (DLCM), for HPE. We first show each bottom-up/top-down inference step of general compositional models is indeed an instantiation of a generalized process we call *spatially local information summarization* (SLIS). As shown in Fig. 2(b), it aggregates information from input score maps[2] on a spatially local support to predict output score maps. In this paper, we exploit CNNs to model this process due to their capability to approximate inference functions via spatially local connections. As a result, DLCMs can learn more sophisticated and realistic compositional patterns within human bodies. To avoid potentially large state spaces, we propose to use state variables to only denote locations and embed the type information into score maps. Specially, we use bone segments to represent a part and supervise its score map in the training phase. This novel representation not only compactly encodes the orientation, scale and shape of a part, but also reduces both computation and space complexities. Fig. 2(c) provides an overview of a DLCM. We evaluate the proposed approach on three HPE benchmarks. With significantly less parameters and lower computational complexities, it outperforms state-of-the-art methods.

In summary, the novelty of this paper is as follows:

- To the best of our knowledge, this is the first attempt to explicitly learn the hierarchical compositionality of visual patterns via deep neural networks. As a result, DLCMs are capable to characterize the complex and realistic compositional relationships among body parts.
- We propose a novel part representation. It encodes the orientation, scale and shape of each part compactly and avoids their potentially large state spaces.
- Compared with prior deep neural networks, *e.g.*, CNNs, designed for HPE, our model has a hierarchical compositional structure and bottom-up/top-down inference stages across multiple semantic levels. We show in the experiments that the compositional nature of DLCMs helps them resolve the ambiguities that appear in bottom-up pose predictions.

2 Related Work

Compositional models. Compositionality has been studied in several lines of vision research [13, 24, 14, 25] and exploited in tasks like HPE [12, 15–19, 26],

[2] Each entry of a score map evaluates the goodness of a part being at a certain state, *e.g.*, location and type.

semantic segmentation [27] and object detection [28]. However, prior compositional models adopt simple and unrealistic relational modeling, *e.g.*, pairwise potentials based on Gaussian distributions. They are incapable to model complex compositional patterns. Our approach attempts to address this difficulty by learning the compositional relationships among body parts via the powerful CNNs. In addition, we exploit a novel part representation to compactly encode the scale, orientation and shape of each part and avoid their potentially large state spaces.

CNN-based HPE. All state-of-the-art HPE systems take CNNs as their main building block [5–7, 9, 29]. Newell *et. al.* [5] introduce a novel *hourglass* module to process and consolidate features across all scales to best capture the various spatial relationships associated with the body. Yang *et. al.* [7] combine CNNs and the expressive deformable mixture of parts [30] to enforce the spatial and appearance consistency among body parts. Hu and Ramanan [29] unroll the inference process of hierarchical rectified Gaussians as bidirectional architectures that also reason with top-down feedback. Instead of predicting body joint positions directly, Sun *et. al.* [31] regress the coordinate shifts between joint pairs to encode their interactions. It is worth noting that none of these methods decomposes entities as hierarchies of meaningful and reusable parts or infers across different semantic levels. Our approach differs from them in that: (1) It has a hierarchical compositional network architecture; (2) CNNs are used to learn the compositional relationships among body parts; (3) Its inference consists of both bottom-up and top-down stages across multiple semantic levels; (4) It exploits a novel part representation to supervise the training of CNNs.

Bone-based part representations. Some prior works [32, 33] use heat maps of limbs between each pair of adjacent joints as supervisions of deep neural networks. Their motivation is that modeling pairs of joints helps capture additional body constraints and correlations. Different with them, our bone-based part representation has (1) a hierarchical compositional structure and (2) multiple semantic levels. It is designed to (1) tightly encode the scale, orientation and shape of a part, (2) avoid exponentially large state spaces for higher-level parts and (3) guide CNNs to learn the compositionality of human bodies.

3 Our Approach

We first make a brief introduction to general compositional models (Sec. 3.1). Their inference steps are generalized as SLIS functions and modeled with CNNs (Sec. 3.2). We then describe our novel bone-based part representation (Sec. 3.3). Finally, the deeply learned compositional models are detailed in Sec. 3.4.

3.1 Compositional models

A compositional model is defined on a hierarchical graph, as shown in Fig. 3. It is characterized by a 4-tuple $(\mathcal{V}, \mathcal{E}, \phi^{and}, \phi^{leaf})$, which specifies its graph structure $(\mathcal{V}, \mathcal{E})$ and potential functions $(\phi^{and}, \phi^{leaf})$. We consider two types of

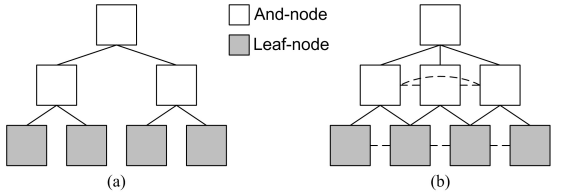

Fig. 3. Example compositional models (a) without and (b) with part sharing and higher-order cliques

nodes[3]: $\mathcal{V} = \mathcal{V}^{and} \cup \mathcal{V}^{leaf}$. And-nodes $\mathcal{V}^{and}$ model the composition of subparts into higher-level parts. Leaf nodes $\mathcal{V}^{leaf}$ model primitives, *i.e.*, the lowest-level parts. We call And-nodes at the highest level as root nodes. $\mathcal{E}$ denotes graph edges. In this section, we first illustrate our idea using the basic compositional model shown in Fig. 3(a), which does not share parts and considers only pairwise relationships, and then extend it to the general one, as shown in Fig. 3(b).

A state variable w_u is associated to each node/part $u \in \mathcal{V}$. For HPE, it can be the position p_u and type t_u of this part: $w_u = \{p_u, t_u\}$. As a motivating example, Yang and Ramanan [30] use types to represent orientations, scales and semantic classes (a straight versus bended arm) of parts.

Let Ω denote the set of all state variables in the model. The probability distribution over Ω is of the following Gibbs form:

$$p(\Omega|\mathbf{I}) = \frac{1}{Z}\exp\{-E(\Omega, \mathbf{I})\} \tag{1}$$

where $\mathbf{I}$ is the input image, $E(\Omega, \mathbf{I})$ is the energy and Z is the partition function. For convenience, we use a score function $S(\Omega)$, defined as the negative energy, to specify the model and omit $\mathbf{I}$. Without part sharing and higher-order potentials, it can be written as:

$$S(\Omega) \equiv -E(\Omega, \mathbf{I}) = \sum_{u \in \mathcal{V}^{leaf}} \phi_u^{leaf}(w_u, \mathbf{I}) + \sum_{u \in \mathcal{V}^{and}} \sum_{v \in ch(u)} \phi_{u,v}^{and}(w_u, w_v) \tag{2}$$

where $ch(u)$ denotes the set of children of node u. The two terms are potential functions corresponding to Leaf and And nodes, respectively. The first term acts like a detector: it determines how likely the primitive modeled by Leaf-node u is present at location p_u and of type t_u. The second term models the state compatibility between a subpart v and its parent u.

Thanks to the tree structure, the optimal states Ω^* for an input image $\mathbf{I}$ can be computed efficiently via dynamic programming. We call this process the *compositional inference*. It is consisted of two stages. In the bottom-up stage,

[3] We do not need Or-nodes [13, 14] here as part variations have been explicitly modeled by the state variables of And-nodes.

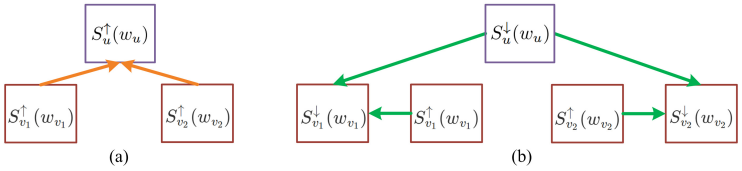

Fig. 4. Illustration of input-output relationships between child and parent score maps in the compositional inference. In this example, node u has two children v_1 and v_2. (a) In the bottom-up stage, the score map of a higher-level part is a function of its children's score maps. (b) In the top-down stage, the score map of a lower-level part is refined by its parent's score map updated one step earlier

the maximum score, *i.e.*, $\max_{\Omega} S(\Omega)$, can be calculated recursively as:

$$\text{(Leaf)}\ S_u^{\uparrow}(w_u) = \phi_u^{leaf}(w_u, \mathbf{I}) \tag{3}$$

$$\text{(And)}\ S_u^{\uparrow}(w_u) = \sum_{v \in ch(u)} \max_{w_v} [\phi_{u,v}^{and}(w_u, w_v) + S_v^{\uparrow}(w_v)] \tag{4}$$

where $S_u^{\uparrow}(w_u)$ is the maximum score of the subgraph formed by node u and all its descendants, with root node u taking state w_u, and is computed recursively by Eq. (4), with boundary conditions provided by Eq. (3). The recursion begins from the Leaf-level and goes up until root nodes are reached. As a function, $S_u^{\uparrow}(w_u)$ assigns each possible state of part u a score. It can also be considered as a tensor/map, each entry of which is indexed by the part's state and valued by the corresponding score. Thus, we also call $S_u^{\uparrow}(w_u)$ the *score map* of part u.

In the top-down stage, we recursively invert Eq. (4) to obtain the optimal states of child nodes that yield the maximum score:

$$\text{(Root)}\quad w_u^* = \text{argmax}_{w_u}\, S_u^{\downarrow}(w_u) \equiv \text{argmax}_{w_u}\, S_u^{\uparrow}(w_u) \tag{5}$$

$$\text{(Non-root)}\ w_v^* = \text{argmax}_{w_v}\, S_v^{\downarrow}(w_v) \equiv \text{argmax}_{w_v} [\phi_{u,v}^{and}(w_u^*, w_v) + S_v^{\uparrow}(w_v)] \tag{6}$$

where node u in Eq. (6) is the unique parent of node v, *i.e.*, $\{u\} = pa(v)$, $S_u^{\uparrow}(w_u)$ and $S_v^{\uparrow}(w_v)$ are acquired from the bottom-up stage, $S_u^{\downarrow}(w_u)$ and $S_v^{\downarrow}(w_v)$ are respectively refined score maps of nodes u and v. Specially, w_u^* and w_v^* are respectively optimal states of parts u and v, and are computed recursively by Eq. (6), with boundary conditions provided by Eq. (5). The recursion begins from root nodes and goes down until the Leaf-level is reached.

3.2 Spatially local information summarization

From Eq. (6), $S_v^{\downarrow}(w_v)$ for non-root nodes is defined as:

$$S_v^{\downarrow}(w_v) = \phi_{u,v}^{and}(w_u^*, w_v) + S_v^{\uparrow}(w_v) \tag{7}$$

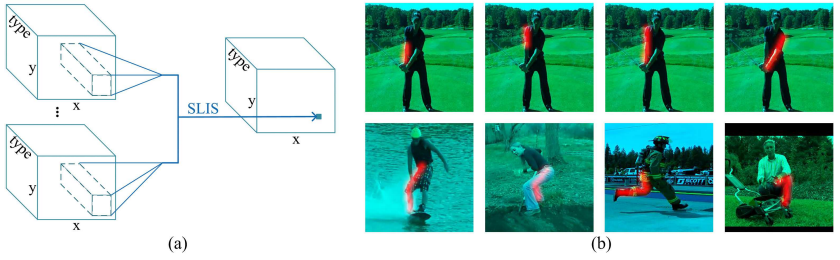

Fig. 5. (a) Illustration of the SLIS function in the compositional inference. Each cube denotes a score map corresponding to a part or subpart. Each entry in the output/right score map is obtained by aggregating information from the input/left score maps on a local spatial support. (b) Illustration of bone-based part representations. First row: the right lower arm, right upper arm, right arm and left arm of a person. Second row: right or left legs of different persons

where $\{u\} = pa(v)$, $w_u^* = \text{argmax}_{w_u} S_u^{\downarrow}(w_u)$. We can write the bottom-up (BU) and top-down (TD) recursive equations, *i.e.*, Eq. (4) and Eq. (7), together as

$$\text{(BU) } S_u^{\uparrow}(w_u) = \sum_{v \in ch(u)} \max_{w_v} [\phi_{u,v}^{and}(w_u, w_v) + S_v^{\uparrow}(w_v)] \tag{8}$$

$$\text{(TD) } S_v^{\downarrow}(w_v) = \sum_{w_u} \phi_{u,v}^{and}(w_u, w_v) \bar{S}_u^{\downarrow}(w_u) + S_v^{\uparrow}(w_v) \tag{9}$$

where $\bar{S}_u^{\downarrow}(w_u)$ is the hard-thresholded version of $S_u^{\downarrow}(w_u)$: $\bar{S}_u^{\downarrow}(w_u)$ equals to 1 if $w_u = w_u^*$ and 0 otherwise. As illustrated in Fig. 4, these two equations intuitively demonstrate how score maps are *propagated* upwards and downwards in the inference process, which finally gives us the globally optimal states, *i.e.*, Ω^*, of the compositional model.

In both equations, there exist summation and/or maximization operations over state variables, *e.g.*, $\sum_{v \in ch(u)} \max_{w_v}$ and $\sum_{w_u}$, as well as between score maps. They can be considered as average and maximum poolings. In the literature of statistical learning [34], pooling means to combine features in a way that preserves task-related information while removing irrelevant details, leads to more compact representations, and better robustness to noise and clutter. In the compositional inference, score maps of some parts are combined to get relevant information about the states of other related parts. This analogy leads us to think of Eqs. (8) and (9) as different kinds of *information summarization*.

Since child and parent parts should not be far apart in practice, it is unnecessary to search them within the whole image [35, 36, 14]. Thus, it is reasonable to constrain their relative displacements to be within a small range: $p_v - p_u \in \mathbb{D}_{uv}$, *e.g.*, $\mathbb{D}_{uv} = [-50, 50] \times [-50, 50]$. For compositional models, this constraint can be enforced by setting $\phi_{u,v}^{and}(w_u, w_v) = 0$ if $p_v - p_u \notin \mathbb{D}_{uv}$. Consequently, for each entry of the score maps on the LHS of Eqs. (8) and (9), only information within a *local spatial region* is summarized on the RHS, as the mapping shown in Fig.

5(a). Note this mapping is also *location-invariant* because the spatial compatibility between parts u and v with types t_u and t_v only depends on their relative locations and is unrelated to their global coordinates in the image space.

Our analysis indicates both recursive equations can be considered as different instantiations of a more generalized process, which aggregates information on a local spatial support and is location-invariant. We call this process *spatially local information summarization* (SLIS) and illustrate it in Fig. 5(a). In the bottom-up stage, the score map of a higher-level part $S_u^{\uparrow}(w_u)$ is an SLIS function of their children's score maps $\{S_v^{\uparrow}(w_v)\}_{v\in ch(u)}$. In the top-down stage, the score map of a lower-level part $S_v^{\downarrow}(w_v)$ is an SLIS function of its parent's score map $S_u^{\downarrow}(w_u)$ as well as its own score map estimated in the bottom-up stage $S_v^{\uparrow}(w_v)$.

Model SLIS functions with CNNs. In this paper, we exploit CNNs to model our SLIS functions for two reasons. First, CNNs aggregate information on a local spatial support using location-invariant parameters. Second, CNNs are known for their capability to approximate inference functions. By learning them from data, we expect the SLIS functions are capable to infer the sophisticated compositional relationships within *real* human bodies. Specifically, we replace Eqs. (8) and (9) with:

$$\text{(BU) } S_u^{\uparrow}(w_u) = \mathbf{c}_u^{\uparrow}\big(\{S_v^{\uparrow}(w_v)\}_{v\in ch(u)}; \Theta_u^{\uparrow}\big) \tag{10}$$

$$\text{(TD) } S_v^{\downarrow}(w_v) = \mathbf{c}_v^{\downarrow}\big(S_u^{\downarrow}(w_u), S_v^{\uparrow}(w_v); \Theta_v^{\downarrow}\big) \tag{11}$$

where $\mathbf{c}_u^{\uparrow}$ and $\mathbf{c}_v^{\downarrow}$ are CNN mappings with $\Theta_u^{\uparrow}$ and $\Theta_v^{\downarrow}$ being their respective collections of convolutional kernels. Since the bottom-up and top-down SLIS functions are different, their corresponding kernels should also be different.

Part sharing and higher-order potentials. We now consider a more general compositional model, as shown in Fig. 3(b). With part sharing and higher-order potentials, the score function is

$$S(\Omega) = \sum_{u\in\mathcal{V}^{leaf}} \phi_u^{leaf}(w_u, \mathbf{I}) + \sum_{u\in\mathcal{V}^{and}} \phi_u^{and}(w_u, \{w_v\}_{v\in ch(u)}) \tag{12}$$

where $\phi_u^{and}(w_u, \{w_v\}_{v\in ch(u)})$ denotes the higher-order potential function measuring the state compatibility among part u and its child parts $\{v : v \in ch(u)\}$.

Due to the existence of loops and child sharing, states of all parts at one level should be estimated/refined jointly from all parts at a lower/higher level. By exploiting the update rules of dynamic programming [25], similar derivations (available in the supplementary material) indicate that we can approximate the SLIS functions as follows:

$$\text{(BU) } \{S_u^{\uparrow}(w_u)\}_{u\in\mathcal{V}^L} = \mathbf{c}_L^{\uparrow}\big(\{S_v^{\uparrow}(w_v)\}_{v\in\mathcal{V}^{L-1}}; \Theta_L^{\uparrow}\big) \tag{13}$$

$$\text{(TD) } \{S_v^{\downarrow}(w_v)\}_{v\in\mathcal{V}^{L-1}} = \mathbf{c}_{L-1}^{\downarrow}\big(\{S_u^{\downarrow}(w_u)\}_{u\in\mathcal{V}^L}, \{S_v^{\uparrow}(w_v)\}_{v\in\mathcal{V}^{L-1}}; \Theta_{L-1}^{\downarrow}\big) \tag{14}$$

where L indexes the semantic level, $\mathcal{V}^L$ denotes the set of nodes at the Lth level, $\Theta_L^{\uparrow}$ and $\Theta_{L-1}^{\downarrow}$ are convolutional kernels. In the bottom-up stage, score maps at a higher level are jointly estimated from all score maps at one level lower. In the top-down stage, score maps at a lower level are jointly refined by all score maps at one level higher as well as their initial estimations in the bottom-up stage.

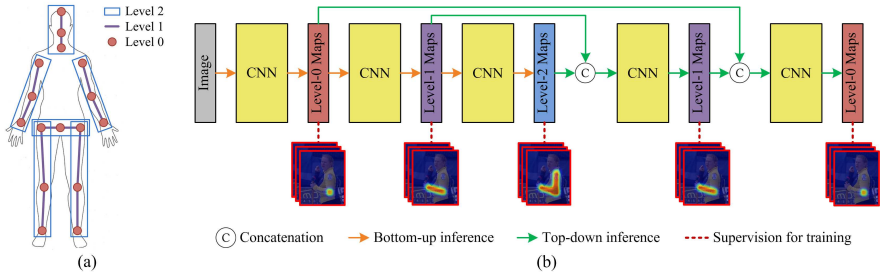

Fig. 6. (a) The compositional structure of a human body used in our experiments. It has three semantic levels, which include 16, 12 and 6 parts, respectively. Assume all the children sharing a common parent are linked to each other. (b) Network architecture of the proposed DLCM. *Maps* in the rectangles are short for score maps

3.3 Bone-based part representation

Another problem with existing compositional models is that the type space for higher-level parts are potentially large. For example, if we have N types for both the left lower leg and left upper leg, there can be $O(N^2)$ types for the whole left leg and $O(N^4)$ types for the composition of left and right legs. As a result, the type dimensions of score maps $S_u^{\uparrow}(w_u)$ and $S_u^{\downarrow}(w_u)$ would be very high, which makes both storage and computation demanding. To address this issue, we propose to embed the type information into score maps and use state variables to only denote locations. As shown in Fig. 5(b), we represent each part with its *bones*, which are generated by putting Gaussian kernels along the part segments. They are then taken as the ground truth of score maps $S_u^{\uparrow}(w_u)$ and $S_u^{\downarrow}(w_u)$ when training neural networks. Specifically, for each point on the line segments of a part, we generate a heat map with a 2D Gaussian (std=1 pixel) centered at it. Then, a single heat map is formed by taking the maximum value from these heat maps at each position.

Our novel part representation has several advantages. First, score maps are now 2-D matrices with no type dimension instead of 3-D tensors. This reduces space and computation complexities in score map predictions. Second, the bones compactly encode orientations, scales and shapes of parts, as shown in Fig. 5(b). We no longer need to discretize them via clustering [12, 15–19, 26]. One weakness of this representation is that the ends of parts are indistinguishable. To solve this problem, we augment score maps of higher-level parts with score maps of their ends[4]. In this way, all important information of parts can be retained.

3.4 Deeply learned compositional model (DLCM)

Motivated by the reasoning above, our Deeply Learned Compositional Model (DLCM) exploits CNNs to learn the compositionality of human bodies for HPE.

[4] In practice, we find repeated ends can be removed without deteriorating performance.

Fig. 6(b) shows an example network based on Eqs. (13) and (14). It has a hierarchical compositional architecture and bottom-up/top-down inference stages. In the bottom-up stage, score maps of target joints are first regressed directly from the image observations, as with existing CNN-based HPE methods. Then, score maps of higher-level parts are recursively estimated from those of their children. In the top-down stage, score maps of lower-level parts are recursively refined using their parents' score maps as well as their own score maps estimated in the bottom-up stage. Similar as [37], a Mean Squared Error (MSE) loss is applied to compare predicted score maps with the ground truth. In this way, we can guide the network to learn the compositional relationships among body parts. Some examples of score maps predicted by our DLCM in the bottom-up and top-down stages can be found in Fig. 8(a).

4 Experiments

4.1 Implementation details

The proposed DLCM is a general framework and can be instantiated with any compositional body structures and CNN modules. In the experiments, we use a similar compositional structure as that in [12] but include higher-order cliques and part sharing. As shown in Fig. 6(a), it has three semantic levels, which include 16, 12 and 6 parts, respectively. Assume all children sharing a common parent are linked to each other. The whole human body is not included here since it has negligible effect on overall performances, while complicating the model.

For two reasons, we exploit the hourglass module [5] to instantiate the CNN blocks in Fig. 6(b). First, the hourglass module extends the fully convolutional network [38] by processing and consolidating features across multiple scales. This enables it to capture the various spatial relationships associated with the input score maps. Second, the eight-stack hourglass network [5], formed by sequentially stacking eight hourglass modules, has achieved state-of-the-art results on several HPE benchmarks. It serves as a suitable baseline to test the effectiveness of the proposed approach. To instantiate a DLCM with three semantic levels, we need five hourglass modules, *i.e.*, the five CNN blocks in Fig. 6(b). Newell *et. al.* [5] add the intermediate features used to predict part score maps back to these predictions via skip connections before they are fed into the next hourglass. We follow this design in our implementation and find it helps reduce overfitting.

Our approach is evaluated on three HPE benchmark datasets of increasing difficulties: FLIC [39], Leeds Sports Poses (LSP) [40] and MPII Human Pose [21]. The FLIC dataset is composed of 5003 images (3987 for training, 1016 for testing) taken from films. The images are annotated on the upper body with most figures facing the camera. The extended LSP dataset consists of 11k training images and 1k testing images from sports activities. As a common practice [6, 41, 9], we train the network by including the MPII training samples. A few joint annotations in the LSP dataset are on the wrong side. We manually correct them. The MPII dataset consists of around 25k images with 40k annotated samples (28k for training, 11k for testing). The images cover a wide range of everyday human

Table 1. Comparisons of PCK@0.2 scores on the FLIC testing set

	Elbow	Wrist	Total
Tompson *et. al.* [42]	93.1	89.0	91.05
Chen&Yuille [43]	95.3	92.4	93.9
Wei *et. al.* [6]	97.6	95.0	96.3
Newell *et. al.* [5]	99.0	97.0	98.0
Ours (3-level DLCM)	**99.5**	**98.5**	**99.0**

Table 2. Comparisons of PCK@0.2 scores on the LSP testing set

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Bulat&Tzimiropoulos [8]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Insafutdinov *et. al.* [44]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Lifshitz *et. al.* [45]	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Yu *et. al.* [46]	87.2	88.2	82.4	76.3	91.4	85.8	78.7	84.3
Chu *et. al.* [9]	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Chen *et. al.* [47]	**98.5**	94.0	89.8	87.5	93.9	94.1	93.0	93.1
Sun *et. al.* [48]	97.9	93.6	89.0	85.8	92.9	91.2	90.5	91.6
Yang *et. al.* [49]	98.3	94.5	92.2	88.9	94.4	95.0	93.7	93.9
Ours (3-level DLCM)	98.3	**95.9**	**93.5**	**90.7**	**95.0**	**96.6**	**95.7**	**95.1**

activities and a great variety of full-body poses. Following [42, 5], 3k samples are taken as a validation set to tune the hyper-parameters.

Each input image is cropped around the target person according to the annotated body position and scale. They are then resized to 256×256 pixels. Data augmentation based on affine transformation [50, 48] is used to reduce overfitting. We implement DLCMs[5] using Torch [51] and optimize them via RMSProp [52] with batch size 16. The learning rate is initialized as 2.5×10^{-4} and then dropped by a factor of 10 after the validation accuracy plateaus. In the testing phase, we run both the original input and a flipped version of a six-scale image pyramid through the network and average the estimated score maps together [49]. The final prediction is the maximum activating location of the score map for a given joint predicted by the last CNN module.

4.2 Evaluation

Metrics. Following previous work, we use the Percentage of Correct Keypoints (PCK) [21] as the evaluation metric. It calculates the percentage of detections that fall within a normalized distance of the ground truth. For LSP and FLIC, the distance is normalized by the torso size, and for MPII, by a fraction of the head size (referred to as PCKh).

Accuracies. Tabs. 1-3 respectively compare the performances of our 3-level DLCM and the most recent state-of-the-art HPE methods on FLIC, LSP and

[5] `http://www.ece.northwestern.edu/~wtt450/project/ECCV18_DLCM.html`

Table 3. Comparisons of PCKh@0.5 scores on the MPII testing set

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Bulat&Tzimiropoulos [8]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Gkioxary *et. al.* [53]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Insafutdinov *et. al.* [44]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Lifshitz *et. al.* [45]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Belagiannis *et. al.* [33]	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Sun *et. al.* [31]	97.5	94.3	87.0	81.2	86.5	78.5	75.4	86.4
Sun *et. al.* [48]	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
Yang *et. al.* [49]	**98.5**	96.7	92.5	**88.7**	91.1	88.6	86.0	92.0
Newell *et. al.* [5]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Ours (3-level DLCM)	98.4	**96.9**	**92.6**	**88.7**	**91.8**	**89.4**	**86.2**	**92.3**

Table 4. Comparisons of parameter and operation numbers

	#parameters	#operations (GFLOPS)
Yang *et. al.* [49] (state-of-the-art)	26.9M	45.9
Newell *et. al.* [5]	23.7M	41.2
Ours (3-level DLCM)	**15.5M**	**33.6**

MPII datasets. Our approach clearly outperforms the eight-stack hourglass network [5], especially on some challenging joints. On the FLIC dataset, it achieves 1.5% improvement on *wrist* and halves the overall error rate (from 2% to 1%). On the MPII dataset, it achieves 2.6%, 2.0%, 1.7%, 1.6% and 1.4% improvements on *ankle*, *knee*, *hip*, *wrist* and *elbow*, respectively. On all three datasets, our approach achieves superior performance to the state-of-the-art methods.

Complexities. Tab. 4 compares the complexities of our 3-level DLCM with the eight-stack hourglass network [5] as well as the current state-of-the-art method [49]. Obviously, using only five hourglass modules instead of eight [5, 49], our model has significantly less parameters and lower computational complexities. Specially, the prior top-performing method [49] on the benchmarks has 74% more parameters and needs 37% more GFLOPS.

Summary. From Tabs. 1-4, we can see that with significantly less parameters and lower computational complexities, the proposed approach has an overall superior performance to the state-of-the-art methods.

4.3 Component analysis

We analyze the effectiveness of each component in DLCMs on MPII validation set. Mean PCKh@0.5 over hard joints, *i.e.*, ankle, knee, hip, wrist and elbow, is used as the evaluation metric. A DLCM with two semantic levels is taken as the basic model. Model (i), $i \in \{1, 2, 3, 4, 5\}$, denotes one of the five variants of the basic model shown in Fig. 7(a).

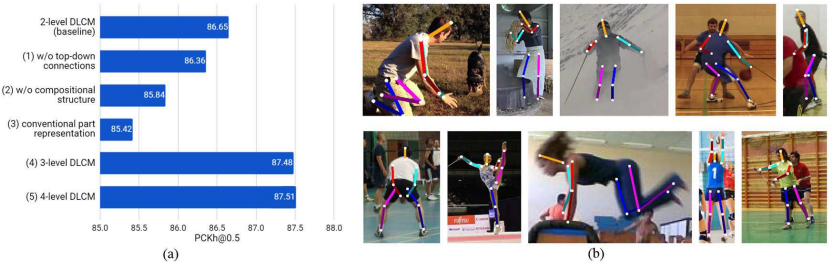

Fig. 7. (a) Component analysis on MPII validation set. See Sec. 4.3 for details. (b) Qualitative results obtained by our approach on the MPII (top row) and LSP (bottom row) testing sets

To see the importance of compositional architectures, we successively remove the top-down lateral connections and compositional part supervisions, which leads to Model (1) and Model (2). Fig. 7(a) indicates that both variants, especially the second one, perform worse than the basic model.

In Model (3), we replace bone-based part representations in the basic model with conventional part representations, *i.e.*, cubes in Fig. 5(a). Following [12], we use K-means to cluster each of the 12 higher-level parts into N types. Since a part sample is assigned to one type, only 1 of its N score map channels is nonzero (with a Gaussian centered at the part location). We have tested $N = 15$ [12] and $N = 30$ and reported the better result. As shown in Fig. 7(a), the novel bone-based part representation significantly outperforms the conventional one.

Finally, we explore whether using more semantic levels in a DLCM can boost its performance. Model (4) is what we have used in Sec. 4.2. Model (5) has 4 semantic levels. The highest-level part is the whole human body. Its ground truth *bone* map is the composition (location-wise maximum) of its children's *bone* maps. Fig. 7(a) shows that the 3-level DLCM performs much better than the 2-level model. However, with 38% more parameters and 27% more GFLOPS, the 4-level DLCM only marginally outperforms the 3-level model.

4.4 Qualitative results

Fig. 7(b) displays some pose estimation results obtained by our approach. Fig. 8(a) visualizes some score maps obtained by our method in the bottom-up (BU) and top-down (TD) inference stages. The evolution of these score maps demonstrates how the learned compositionality helps resolve the low-level ambiguities that appear in high-level pose estimations. The uncertain bottom-up estimations of the left ankle, right ankle and right elbow respectively in the first, second and fifth examples are resolved by the first-level compositions. In some more challenging cases, one level of composition is not enough to resolve the ambiguities, *e.g.*, the bottom-up predictions of the left lower arm in the third example and the left lower leg in the fourth example. Thanks to the hierarchical compositionality, their uncertainties can be reduced by the higher-level relational models. Fig.

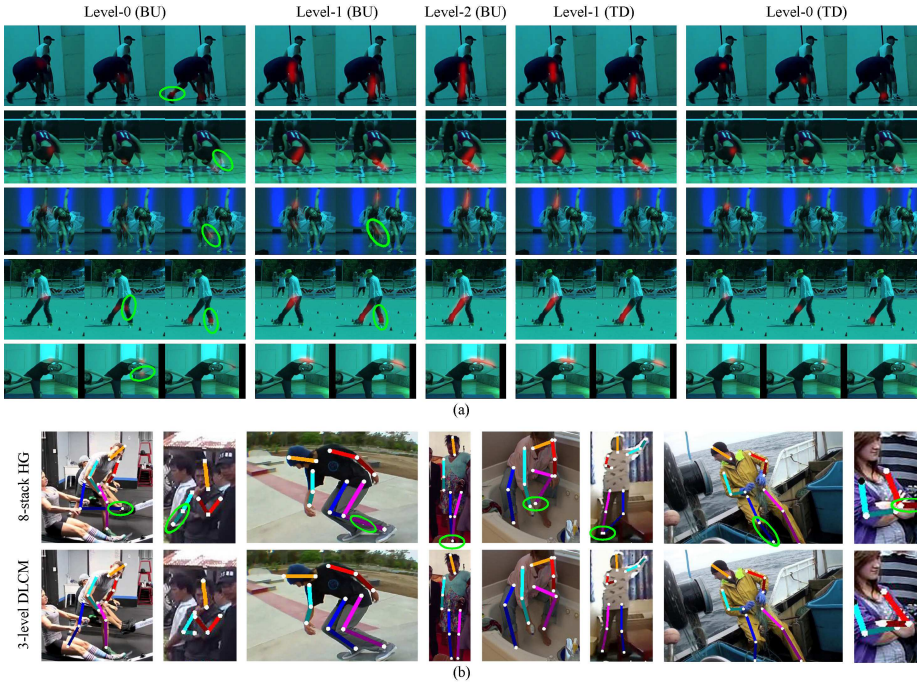

Fig. 8. (a) Score maps obtained by our method on some unseen images in the bottom-up (BU) and top-down (TD) inference stages. The five columns correspond to the five inference steps in Fig. 6(b). Due to space limit, only score maps corresponding to one of the six level-2 parts are displayed for the example at each row. From top to bottom, the level-2 parts are left leg, right leg, left arm, left leg and right arm, respectively. Within each sub-figure, parts of the same level are ordered by their distances to the body center. (b) Some examples showing that a 3-level DLCM (bottom row) is able to resolve the ambiguities that appear in bottom-up pose predictions of an 8-stack hourglass network (top row). Wrong part localizations are highlighted by green ellipses

8(b) shows that our DLCM can resolve the ambiguities that appear in bottom-up pose predictions of an 8-stack hourglass network.

5 Conclusion

This paper exploits deep neural networks to learn the complex compositional patterns within human bodies for pose estimation. We also propose a novel bone-based part representation to avoid potentially large state spaces for higher-level parts. Experiments demonstrate the effectiveness and efficiency of our approach.

Acknowledgement. This work was supported in part by National Science Foundation grant IIS-1217302, IIS-1619078, and the Army Research Office ARO W911NF-16-1-0138.

References

1. Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A.: 3d human pose estimation: A review of the literature and analysis of covariates. Computer Vision and Image Understanding **152** (2016) 1–20
2. Fukushima, K., Miyake, S.: Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and cooperation in neural nets. Springer (1982) 267–285
3. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
4. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553) (2015) 436
5. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision, Springer (2016) 483–499
6. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4724–4732
7. Yang, W., Ouyang, W., Li, H., Wang, X.: End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3073–3082
8. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: European Conference on Computer Vision, Springer (2016) 717–732
9. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 5669–5678
10. Geman, S., Potter, D.F., Chi, Z.: Composition systems. Quarterly of Applied Mathematics **60**(4) (2002) 707–736
11. Bienenstock, E., Geman, S., Potter, D.: Compositionality, mdl priors, and object recognition. In: Advances in neural information processing systems. (1997) 838–844
12. Tian, Y., Zitnick, C.L., Narasimhan, S.G.: Exploring the spatial hierarchy of mixture models for human pose estimation. In: European Conference on Computer Vision, Springer (2012) 256–269
13. Zhu, S.C., Mumford, D., et al.: A stochastic grammar of images. Volume 2. Now Publishers, Inc. (2007)
14. Zhu, L.L., Chen, Y., Yuille, A.: Recursive compositional models for vision: Description and review of recent work. Journal of Mathematical Imaging and Vision **41**(1-2) (2011) 122
15. Wang, Y., Tran, D., Liao, Z.: Learning hierarchical poselets for human parsing. In: IEEE Conference on Computer Vision and Pattern Recognition. (2011) 1705–1712
16. Rothrock, B., Park, S., Zhu, S.C.: Integrating grammar and segmentation for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2013) 3214–3221
17. Sun, M., Savarese, S.: Articulated part-based model for joint object detection and pose estimation. In: IEEE International Conference on Computer Vision. (2011) 723–730
18. Park, S., Zhu, S.C.: Attributed grammars for joint estimation of human attributes, part and pose. In: IEEE International Conference on Computer Vision. (2015) 2372–2380

19. Park, S., Nie, B.X., Zhu, S.C.: Attribute and-or grammar for joint parsing of human pose, parts and attributes. IEEE transactions on pattern analysis and machine intelligence **40**(7) (2018) 1555–1569
20. Felzenszwalb, P.F., Huttenlocher, D.P.: Distance transforms of sampled functions. Theory of computing **8**(1) (2012) 415–428
21. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on computer Vision and Pattern Recognition. (2014) 3686–3693
22. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: IEEE Conference on Computer vision and pattern recognition. (2011) 1465–1472
23. Tran, D., Forsyth, D.: Improved human parsing with a full relational model. In: European Conference on Computer Vision, Springer (2010) 227–240
24. Jin, Y., Geman, S.: Context and hierarchy in a probabilistic image model. In: IEEE Conference on Computer Vision and Pattern Recognition. (2006) 2145–2152
25. Tang, W., Yu, P., Zhou, J., Wu, Y.: Towards a unified compositional model for visual pattern modeling. In: IEEE International Conference on Computer Vision. (2017) 2803–2812
26. Duan, K., Batra, D., Crandall, D.J.: A multi-layer composite model for human pose estimation. In: British Machine Vision Conference. (2012)
27. Wang, J., Yuille, A.L.: Semantic part segmentation using compositional model combining shape and appearance. In: IEEE conference on computer vision and pattern recognition. (2015) 1788–1797
28. Zhu, L., Chen, Y., Torralba, A., Freeman, W., Yuille, A.: Part and appearance sharing: Recursive compositional models for multi-view. In: IEEE Conference on Computer Vision and Pattern Recognition. (2010) 1919–1926
29. Hu, P., Ramanan, D.: Bottom-up and top-down reasoning with hierarchical rectified gaussians. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016) 5600–5609
30. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: IEEE Conference o Computer Vision and Pattern Recognitionn. (2011) 1385–1392
31. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: IEEE International Conference on Computer Vision. (2017) 2621–2630
32. Ai, B., Zhou, Y., Yu, Y., Du, S.: Human pose estimation using deep structure guided learning. In: IEEE Winter Conference on Applications of Computer Vision. (2017) 1224–1231
33. Belagiannis, V., Zisserman, A.: Recurrent human pose estimation. In: IEEE International Conference on Automatic Face Gesture Recognition. (2017) 468–475
34. Boureau, Y.L., Ponce, J., LeCun, Y.: A theoretical analysis of feature pooling in visual recognition. In: International Conference on Machine Learning. (2010) 111–118
35. Wan, L., Eigen, D., Fergus, R.: End-to-end integration of a convolution network, deformable parts model and non-maximum suppression. In: IEEE Conference on Computer Vision and Pattern Recognition. (2015) 851–859
36. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**(9) (2010) 1627–1645
37. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in neural information processing systems. (2014) 1799–1807

38. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE conference on computer vision and pattern recognition. (2015) 3431–3440
39. Sapp, B., Taskar, B.: Modec: Multimodal decomposable models for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2013) 3674–3681
40. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: British Machine Vision Conference. (2010)
41. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4929–4937
42. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition. (2015) 648–656
43. Chen, X., Yuille, A.L.: Articulated pose estimation by a graphical model with image dependent pairwise relations. In: Advances in neural information processing systems. (2014) 1736–1744
44. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision, Springer (2016) 34–50
45. Lifshitz, I., Fetaya, E., Ullman, S.: Human pose estimation using deep consensus voting. In: European Conference on Computer Vision, Springer (2016) 246–260
46. Yu, X., Zhou, F., Chandraker, M.: Deep deformation network for object landmark localization. In: European Conference on Computer Vision, Springer (2016) 52–70
47. Chen, Y., Shen, C., Wei, X.S., Liu, L., Yang, J.: Adversarial posenet: A structure-aware convolutional network for human pose estimation. In: IEEE International Conference on Computer Vision. (2017) 1221–1230
48. Sun, K., Lan, C., Xing, J., Zeng, W., Liu, D., Wang, J.: Human pose estimation using global and local normalization. In: IEEE International Conference on Computer Vision. (2017) 5600–5608
49. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: The IEEE International Conference on Computer Vision. (2017) 1290–1299
50. Forsyth, D.A., Ponce, J.: Computer Vision: A Modern Approach. Prentice Hall Professional Technical Reference (2002)
51. Collobert, R., Kavukcuoglu, K., Farabet, C.: Torch7: A matlab-like environment for machine learning. In: NIPS Workshop. (2011)
52. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning **4**(2) (2012) 26–31
53. Gkioxari, G., Toshev, A., Jaitly, N.: Chained predictions using convolutional neural networks. In: European Conference on Computer Vision, Springer (2016) 728–743