Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation

Guanghan Ning ⓘ , *Student Member, IEEE*, Zhi Zhang ⓘ , *Student Member, IEEE*, and Zhiquan He

Abstract—Human pose estimation using deep neural networks aims to map input images with large variations into multiple body keypoints, which must satisfy a set of geometric constraints and interdependence imposed by the human body model. This is a very challenging nonlinear manifold learning process in a very high dimensional feature space. We believe that the deep neural network, which is inherently an algebraic computation system, is not the most efficient way to capture highly sophisticated human knowledge, for example those highly coupled geometric characteristics and interdependence between keypoints in human poses. In this work, we propose to explore how external knowledge can be effectively represented and injected into the deep neural networks to guide its training process using learned projections that impose proper prior. Specifically, we use the stacked hourglass design and inception-resnet module to construct a fractal network to regress human pose images into heatmaps with no explicit graphical modeling. We encode external knowledge with visual features, which are able to characterize the constraints of human body models and evaluate the fitness of intermediate network output. We then inject these external features into the neural network using a projection matrix learned using an auxiliary cost function. The effectiveness of the proposed inception-resnet module and the benefit in guided learning with knowledge projection is evaluated on two widely used human pose estimation benchmarks. Our approach achieves state-of-the-art performance on both datasets.

Index Terms—Human pose estimation, fractal networks, knowledge-guided learning.

I. INTRODUCTION

THE task of human pose estimation is to determine the precise pixel locations of body keypoints from a single input image [1]–[7]. Closely-related tasks include 3D human pose estimation [8] and human pose estimation in videos [9], [10]. Human pose estimation is very important for many high-level computer vision tasks, including action and activity recognition [11]–[13], semantic content retrieval [14], human-computer in-

G. Ning and Z. Zhang are with the Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211 USA (e-mail: gnxr9@mail.missouri.edu; zzbhf@mail.missouri.edu).

Z. He is with the College of Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: zhiquan@szu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

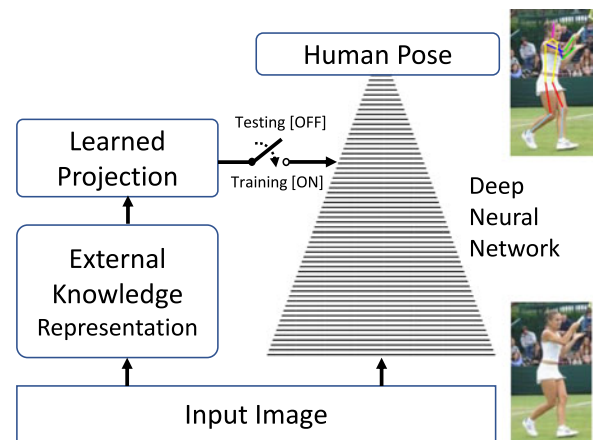Digital Object Identifier 10.1109/TMM.2017.2762010

Fig. 1. *Knowledge projection for guided learning*. We encode external knowledge visual features, which characterize the constraints of human body models and then inject these external features into the neural network using a projection matrix learned using an auxiliary cost function, which is removed during testing, therefore not increasing network complexity.

teraction, motion capture [15], and animation. Estimating human poses from still images is a challenging task. An effective human pose estimation system must be able to handle large pose variations, changes in clothing and lighting conditions, severe body deformations, heavy body occlusions [16]–[18]. A key question for addressing these problems is how to extract strong low and mid-level appearance features capturing discriminative as well as relevant contextual information and how to model complex part relationships allowing for effective yet efficient pose inference. Traditional methods for pose estimation are mostly based on Pictorial Structure (PS) models [19]–[24], which model the spatial relations of rigid body parts using a tree model. A major drawback of such models is the need to hand-design the structure of the model in order to capture important problem-specific dependencies amongst the different output variables and at the same time allow for tractable inference.

With Convolutional Neural Networks (ConvNets) and many assistive methods such as batch normalization [25], resnet [26], and inception design [27], [28], human pose estimation has recently achieved significant progress. Even though deep neural networks are capable of fitting large training data through extensive training, the network often needs to be constructed deeper and wider to gain enough representation power [29]. As the network becomes more complex, the learning and training processing become more sophisticated and challenging
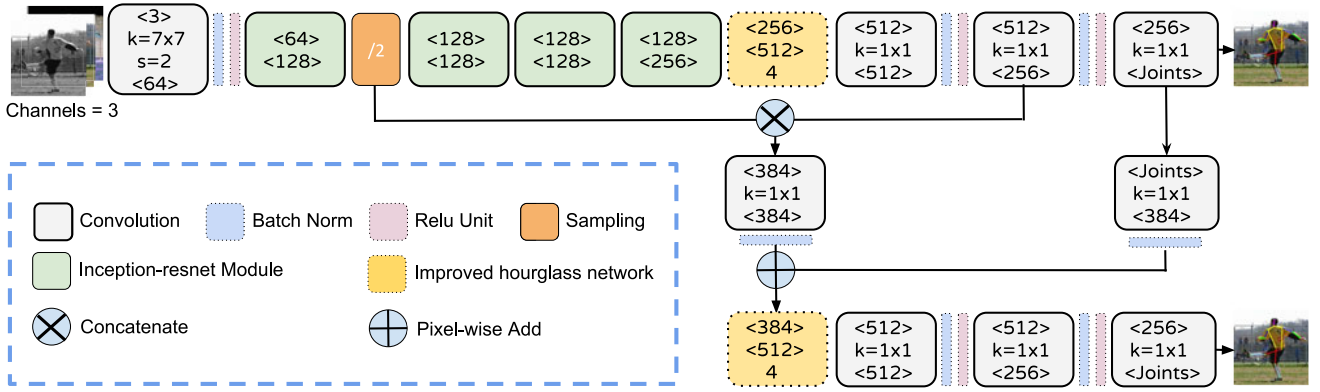
Fig. 2. *Overview of a fractal network*. The network is fractal in that it reflects the concurrence of inception and residual design at both the highest and lowest (inception-resnet module) levels of abstraction. At the top level, images of size 256×256 are downsampled into the resolution of 64×64. Subsequently, inputs and outputs of all modules are of size 64×64, including the output heatmaps. The numbers within brackets in each module denote the number of input and output channels, respectively.

[30], especially for those applications with complicated loss functions.

Human pose estimation using deep neural networks requires us to map the input images with large variations into multiple body keypoints which must satisfy a set of geometric constraints and interdependence imposed by the human body model. This is a very challenging nonlinear manifold learning process in a very high dimensional feature space. We believe that the deep neural network, which is inherently an algebraic computation system, is not the most efficient way to capture highly sophisticated human knowledge, for example those highly coupled geometric characteristics and interdependence between keypoints in human poses.

In this work, we propose to explore how external knowledge can be effectively represented and injected into the deep neural networks to guide its training process using learned projections for more accurate and robust human pose estimation. Specifically, as illustrated in Fig. 2, we use inception-resnet modules and the stacked hourglass structure to construct a fractal network to regress human pose images into heatmaps with no explicit graphical modeling. We encode external knowledge with visual features which characterize the constraints of human body models and evaluate the fitness of intermediate network output. We then inject these external features into the neural network using a projection matrix learned using an auxiliary cost function. The guidance from the external knowledge is only used during the training process, and is turned off during network inference for human pose estimation. The benefit of external knowledge is to guide the training of the neural network. Its effect is implicitly imposed on the tuning of the weight parameters, instead of explicit feature representation of the network. The injected features for pairs of limbs impose a strong prior during the training, preventing human part keypoint from connecting to noises, e.g., keypoint from other people in the background that is not cropped out for the target person.

The major contributions of this work are summarized as follows: (1) we develop a new framework to represent and project human knowledge to guide the training of deep neu-

ral networks for human pose estimation. This external knowledge project framework is generic and can be extended to other learning and training applications and deep neural network design. (2) We propose an efficient network structure, called *fractal networks*, for human pose estimation to capture the multi-scale interdependence between body joints in the pose model. This fractal network uses an *inception-resnet module* as the building block.

The rest of the paper is organized as follows. In Section II, we provide a brief review of recent works on human pose estimation. Section III introduces the concept of knowledge guided learning, the structure of fractal network, and the design of inception-resnet module. Section V presents our experimental results. Section VI concludes our paper.

II. RELATED WORK

A. Structured Prediction and Graphical Models

Prior to the advent of neural networks most previous work on pose estimation was based on pictorial structures [31] which model the human body as a collection of rigid templates and a set of pairwise potentials taking the form of a tree structure, thus allowing for efficient and exact inference at test time. High level knowledge of the human body is exploited by modeling humans with body parts that are connected via a skeleton structure. Pictorial structure models [31], [32] capture the spatial relation of rigid body parts using a tree model. A pre-defined kinematic body model is often used which assumes that each body part is independent of all the others except for the ones it is attached to. A major drawback of such models is the need to hand-design the structure of the model in order to capture important problem-specific dependencies amongst the different output variables and at the same time allow for tractable inference.

Recent work includes sophisticated extensions like mixture, hierarchical, multimodal and strong appearance models [19], [20], [22], [33], [34], non-tree models [23], [24] as well as cascaded/sequential prediction models such as pose machines [35]. While in [31] each limb is represented by a single template that

is parameterized by location, orientation, shape parameters, and an appearance model, Yang and Ramanan [33] propose mixtures of part templates where body part is represented by a set of deformable part templates. Although this approach performs well in comparison to classical pictorial structure models for human pose estimation, it has some limitations. For instance, the used scanning-window templates trained with linear SVMs and HOG features [36] are very sensitive to noise [37]. Hierarchical models [21], [22] represent the relationships between parts at different scales and sizes in a hierarchical tree structure. The underlying assumption of these models is that larger parts (that correspond to full limbs instead of joints) can often have discriminative image structure that can be easier to detect and consequently help reason about the location of smaller, harder-to-detect parts. On the other hand, there are non-tree models [23], [24] to incorporate interactions that introduce loops to augment the tree structure with additional edges that capture symmetry, occlusion and long-range relationships. These methods usually have to rely on approximate inference during both learning and at test time.

B. Deep Neural Networks for Human Pose Regression

ConvNets have been shown to produce remarkable performance for a variety of difficult computer vision tasks including detection [38], [39], recognition [26], [40], and semantic segmentation [41]. A key feature of these approaches is that they integrate non-linear hierarchical feature extraction with the classification or regression task in hand being also able to capitalize on very large data sets that are now readily available.

Since the work of *DeepPose* by Toshev *et al.* [16], research on human pose estimation has shifted from traditional approaches to deep neural networks (DNN) due to their superior performance. Within the context of human pose estimation, it is natural to formulate the problem as a regression one in which CNN features are regressed in order to provide joint predictions of the body parts [10], [16], [42]. For non-visible parts, learning the complex mapping from occluded part appearances to part locations is hard and the network has to rely on contextual information provided by other visible parts to infer the occluded part locations. DeepPose uses a deep neural network to directly regress the coordinates of body joints. Tompson *et al.* [17] argued that it is more efficient to use DNNs to regress heatmap images at multiple scales. While body models are not necessary for effective part localization, constraints between parts allow us to assemble independent detections into a body configuration. Detection-based methods are relying on powerful CNN-based part detectors which are then combined using a graphical model [17], [43] or refined using regression [44], [45]. Regression-based methods try to learn a mapping from the input image and CNN features to part locations. [43] achieved promising results by combining CNN-based body part detectors with a body model [33].

Human pose estimation methods using deep neural networks have proven their significant advantages over traditional approaches. However, deeper and wider networks are often required to improve the feature representation power, which in
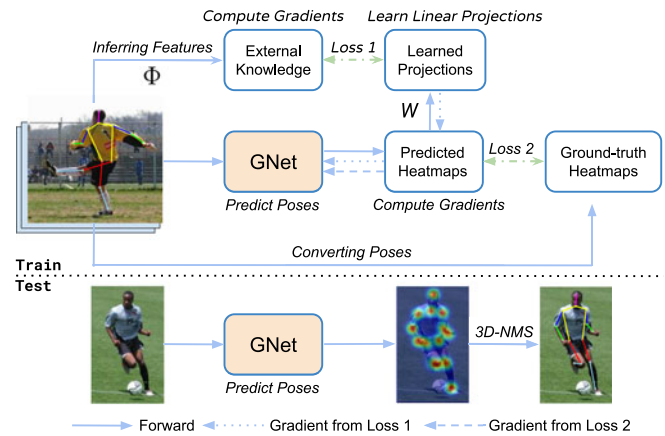

Fig. 3. *Framework of our proposed Guided Network (GNet).* The projected knowledge affects the gradients propagated back to convolutional layers, but they are not part of the network during deployment or test. By enforcing constraints with external knowledge injection, high-level information of long-range dependencies between image and multipart cues that is hard to capture with implicit learning can be better learned under the guidance of mid-level knowledge projection.

turn leads to increased difficulty in training the neural networks. Recently, residual learning [26] has been used to significantly improve the performance of human pose estimation [18], [46]. It was used for part detection in the system of [46]. *stacked hourglass network* of [18] elegantly extends fully convolutional networks [47] and deconvolution nets [48] with residual learning.

Intermediate supervision [49], recursive prediction [50], and inception design [27], [28] are among other successful techniques that have been applied by recent methods for human pose estimation. Recently, researchers recognize that successive predictions can boost the performance of pose estimation, where parts are sequentially refined [16], [35], [51], [52]. In these models an initial prediction is made of all the parts; in subsequent steps, all part predictions are refined based on the image and earlier part predictions. Tompson *et al.* [44] use a cascade of networks for refined predictions to achieve significantly improved precision in joint localization. Carreira *et al.* [52] introduce a so-called *Iterative Error Feedback* scheme, where a set of predictions is included in the input, and each pass through the network further refines these predictions. Their method requires multi-stage training and the weights are shared between iterations. Recently, adding supervision to intermediate layers of deep networks is also explored to assist the training process [27], [53]. Methods in [18], [46], [50], [51] use intermediate supervision to add auxiliary supervision branches in the network to assist the training process for human pose estimation. These approaches all employ the inception design by concatenating heatmaps from different stages or abstract levels as the input for the next layers.

One direction for further improvement of human pose estimation is to design convolutional networks that can produce robust visual features. Multi-scale processing by repetitive down-sampling and up-sampling has been introduced in *Stacked Hourglass Networks* [18]. Another approach to improve human pose
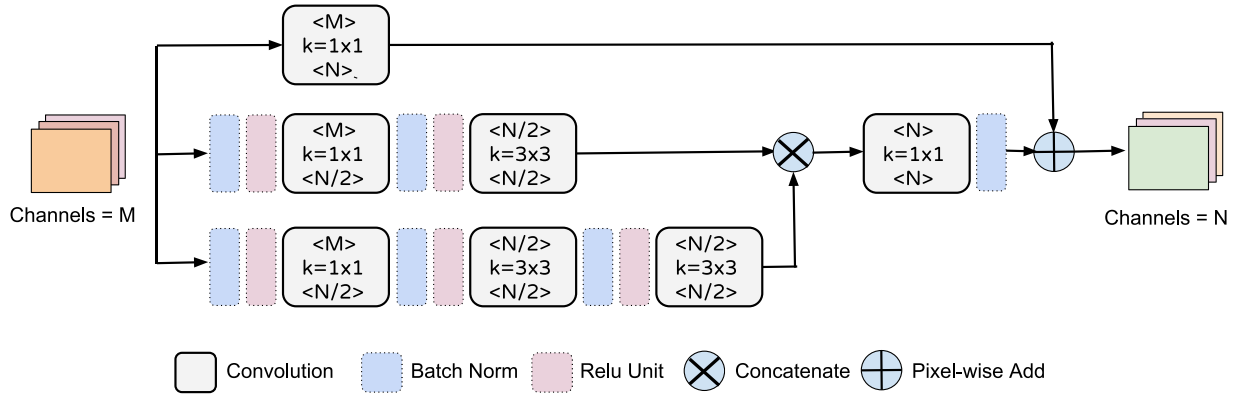
Fig. 4. *Basic module: Inception-resnet*. Convolution layers are padded such that the resolution of output is the same as that of the input. The benefit of this module is that the input and output are of uniform resolution, while the depth of channels can be changed. The function of this basic module is to interpret the input information from one form to another, extracting features for another abstraction level with little loss of information quantity.

estimation performance is to use explicit part-based models [23], [24], [33] or implicitly encode configuration model using its contexts [54]. These methods involve additional sub-networks to detect parts, which increases the overall complexity. In this work, we leverage these ideas and approaches. We propose a fractal network structure using inception-resnet as building blocks to explore the multi-scale interdependence nature of human pose configuration and to capture these characteristics across different scales and resolutions. The network is fractal in that it reflects the co-occurrance of inception and residual design at both the highest and lowest levels of abstractions.

C. Transfer Learning and Guided Training

Nevertheless, training such deep networks has proven to be challenging [55]. Significant efforts has been devoted to alleviate this problem. For instance, there has been another line of work in which a student network is trained from scratch to mimic the behavior of a much larger teacher network. Staring from the work by Bucila *et al.* [56] and Hinton *et al.* [57], Knowledge Distillation (KD) has gained a lot of research interest. In this paper, we consider a unique setting of the problem. Instead of transferring knowledge from teacher networks into a student network, we propose an external knowledge representation and projection framework to guide the training process of our deep neural network for human pose estimation. Specifically, we inject hand-designed features that are inferred from ground truth as external knowledge to aid the training of a highly complex network with deep structure and multiple loss functions. Inspired by [58], which proposed a locality principle to learn task-specific feature mapping for shape regression, we project the external knowledge with a learned feature mapping. The procedure involves domain adaptation and model training simultaneously. Since external knowledge is inferred from the ground truth, it is inherently more reliable and effective than the outputs from a teacher network.

III. PROPOSED METHOD

A. Network Structure and Design

Human pose estimation methods using hand-crafted features or graphical structure models based on human knowledge lack the flexibility in learning and the potential to achieve great representation power. On the other hand, pure data-driven neural networks may not be able to capture sophisticated knowledge involved in human pose estimation. In this work, we propose to represent and inject external human knowledge to guide the learning of deep neural networks (DNN), as illustrated in Fig. 1. Our major idea is that, by enforcing constraints and guidance with external knowledge injection, high-level information of long-range dependencies between image and multi-part cues, which are hard to capture with implicit learning, can be better learned under the guidance of mid-level knowledge projection. As shown in Fig. 3, the projected knowledge affects the gradients propagated back to convolutional layers during training, but they are not part of the network during test.

We borrow the ideas from inception-residual networks [59] and propose to construct a basic inception-resnet module in replacement of convolutional layers for more robust feature representation. Hourglass network is first introduced in [18] where features are processed across all scales by repetitve downsampling and up-sampling and then consolidated to best capture various spatial relationships associated with the body. We introduce a modified version of the hourglass network with the proposed inception-resnet module. As shown in Fig. 2, we use inception-resnet modules and improved hourglass sub-networks to construct a fractal network to regress human pose images into heatmaps with no explicit graphical modeling. The network is fractal in that it has the same network configuration at all levels of analysis and abstractions. This fractal network is designed to capture the multi-scale interdependence nature of human pose configuration and to represent these characteristics across different scales and resolutions. In the inception network, we perform channel-wise concatenation of two tensors from different sources. This enforces the information represented by the features stored in these tensors to be complementary to each other. It encourages and directs these two sources to work on different concepts to produce a more robust union representation [26]. In the Resnet model, we perform pixel-wise addition of two tensors with the same number of channels. From our experiments, we find that this network design allows us to train the network more effectively, since it enforces two separate tensors to be simultaneously accurate in order to render the expected outputs.
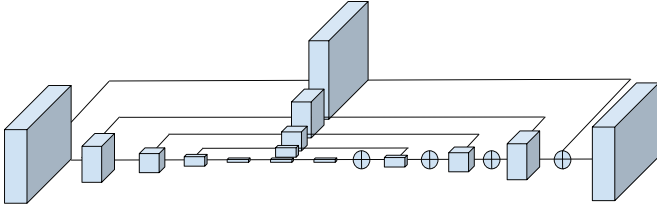
Fig. 5. An illustration of hourglass design proposed in [18]. Pixelwise addition fuses the information from two branches while keeping the input and output resolution uniform. The illustration gives an example of a four-level hourglass.

B. Fractal Network With Inception-Resnet Modules

Our motivation in the fractal network design is that, we need the network to focus on various scales across human parts, and at each scale, the network should also have an overall understanding of this receptive field. At higher levels, the network captures dependencies among various human parts. At lower levels, we use the same fractal design to capture regional dependencies. It is essential to capture local dependencies in addition to local appearances. Because at a certain high-level scale, the receptive field may involve a human part as well as noises from other parts. These adjacent parts may be from the same or other persons. Therefore, local dependencies are helpful in providing more reliable features to higher-level networks.

The construction of inception-resnet module is shown in Fig. 4. Based on the hourglass design proposed in [18] shown in Fig. 5, an improved version of hourglass network is developed in this work as a mid-level sub-network which also uses inception-resnet modules as the basic units, as illustrated in Fig. 6. To combine the advantages of both inception and resnet design, we introduce the inception-resnet module as the basic building block to analyze local fields, while using an improved hourglass network to capture the global information of different parts.

At the bottom level, we propose to use inception-resnet module as the basic structure unit of the network. It consists of convolutional layers, batch norm layers and relu units, with channel-wise concatenation and pixel-wise additions. Convolution layers are padded such that the resolution of output is the same as that of the input. Although the concatenation of two branches maintains different levels of information, the concatenated features across different channels need to be transformed and normalized by the subsequent convolutional layers. In the proposed inception-resnet module, the concatenation layer is followed by another convolutional layer with 1×1 kernels. The benefit of this module is that the input and output have the same resolution while the depth of channels can be flexible.

At the sub-network level, we implement the recursive hourglass for 4 levels as shown in Fig. 5. In other words, it will process the image at four scales. The hourglass network is nested in itself. The first level of hourglass in our network is an inception-resnet module. As illustrated in Fig. 6, we also borrow the idea of hourglass design by down-sampling and then up-sampling the data while using inception-resnet as the basic building block. Pixel-wise addition fuses the information from two branches while keeping the input and output resolution the same.

At the top fractal network level, images of size 256×256 are down-sampled into the resolution of 64×64. In this way, inputs and outputs of all modules are of size 64×64, including the output heatmaps. The network captures and consolidates information across all scales of the image.

C. External Knowledge Representation

The fractal network is used to boost the data representation power of the deep neural network for human pose estimation. As the network grows deeper and more complicated, it requires careful design of the training process. Furthermore, we recognize that the deep neural network is inherently an algebraic computing system, which might not be the most efficient way to capture the highly sophisticated human knowledge during pose estimation, for example those highly coupled geometric constraints and interdependence between body joints. To address these issues, in this work, we propose to encode and inject external knowledge into the fractal network to guide the training process of the network using learned projections.

In this work, we propose to inject the geometric representation of knowledge into the heatmap layer of the network. Since the heatmaps to be predicted are correlated to each other as they share parameters on former layers, the constraint on one heatmap affects the parameters of these layers and therefore having an impact on the training of other heatmaps. We observe that intermediate layers in our network are low and mid-level visual features; higher-level semantic features are hard to locate and explicitly interpret. It is more effective to apply external knowledge and constraints upon to the predicted heatmaps. During the training process, the external knowledge and its visual representations are projected into the background and keypoint heatmaps using a projection matrix. We find that this type of knowledge-guided learning inherently enforces long-range dependencies and configurations among body joints, while preserving the flexibility of visual representation of the network. In the following, we explain the proposed method in more detail.

During the training process, the external knowledge representation module illustrated in Fig. 1 has access to the original training sample image and its ground-truth joint locations. Specifically, during the feature mapping process which is denoted as Φ, we perform the *Hough Transform* on each line traversing two separate joints denoted as (u_i, v_i) and (u_j, v_j). In the Hough space, each line is represented by a coordinate (θ, ρ).

$$\begin{cases} \theta = \arctan\left(\dfrac{x_i - x_j}{y_j - y_i}\right) \\ \rho = x_j \times \cos(\theta) + y_j \times \sin(\theta) \end{cases} \tag{1}$$

In order to represent the information in a more robust manner, we convert the coordinates into a normalized vector representation. To incorporate the inherent learning of geometrical features such as angles and distance, we also inject the joint locations alongside each line. Based on the visibility of each joint, the line traversing it is encoded with the number of visible joints.

In addition to encoding geometric features, we encode image descriptors such as *Histogram of Gradients (HOG)* [36] along the line between each pair of adjacent joints in order to capture
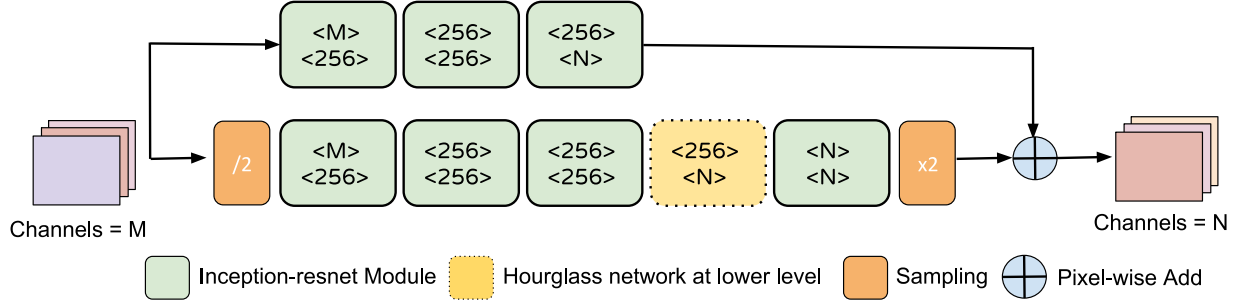
Fig. 6. *Improved hourglass subnetwork*. While using inception-resnet as proposed common building block, we borrow the idea of hourglass design by downsampling and then upsampling the dataflow in one branch, maintaining the resolution of the other branch. The lowest level of the recursive hourglass in our network is an inception-resnet module.
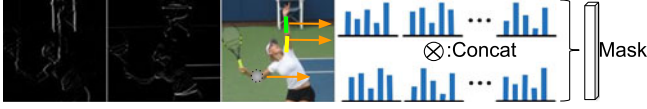

Fig. 7. We encode image descriptors around each pair of adjacent joints in order to capture visual features to compensate spatial dependencies. The features are concatenated and normalized for the external knowledge representation.

visual features to compensate spatial dependencies. While we preserve the power of deep convolutional features that automatically learn visual semantics, we use hand-crafted features as guidance of the learning process by enforcing a strong prior during the training of the neural network. We notice that body joints may connect to those from a nearby person, even though the ground truth joints are not self-occluded or object-occluded. We believe HOG features are effective in characterizing edges and therefore distinguishing real and false limbs which have strong edges. The injected features for pairs of limbs impose a strong prior during the training, preventing body part keypoint from connecting to noises, e.g., keypoints from other people in the background, which is helpful for learning body part inter-dependency. As illustrated in Fig. 7, the features are concatenated and normalized for the external knowledge representation. For self-occluded and object-occluded joints, we mask corresponding features with zeros. Specifically, we follow the traditional HOG feature extraction scheme, applying filters $D_x = [-1 \quad 0 \quad 1]$ and $D_y = [1 \quad 0 \quad -1]^T$ horizontally and vertically to generate gradient maps I_x and I_y. Instead of scanning a window for blocks and cells over the image which is done in traditional ways, we locate limbs based on meta-data from the training set and extract histogram of gradients for such regions. The magnitude and orientation of the gradient are computed by:

$$|G| = \sqrt{I_x^2 + I_y^2} \qquad (2)$$

and

$$\varphi = \arctan \frac{I_y}{I_x} \qquad (3)$$

We use 8 bins for the pooling, followed by block normalization (L2-norm) to mitigate the effect of unbalanced area of regions:

$$f = \frac{v}{\sqrt{||v||_2^2 + e^2}} \qquad (4)$$

Where e is a very small number.

D. Knowledge Projection Into the Deep Neural Network

To translate the external knowledge into a higher-dimensional space, we use two fully-connected (FC) layers and three convolutional layers between the projection representation and the injected knowledge to learn linear projection W, which will be removed during testing. We inject external features as knowledge K via global feature mapping function Φ and learn a global linear projection W by minimizing the loss from the knowledge projection layer:

$$\mathcal{L}_{KP} = ||K - W \times H_J||_2^2 + \beta \times ||W||_2^2 \qquad (5)$$

where the first term is the regression target, the second term is a $L2$ regularization on W, and β controls the regularization strength. Regularization is necessary because the dimensionality of the features is very high. Since the objective function is quadratic with respect to W, we can always reach its global optimum [58].

Specifically, we enforce two loss functions, one for the injected geometric features and one for limb-wise edge features. (1) The ground truth heatmap is convolved by 1×1 kernels, outputting 8 channels of maps. It is padded such that the resolution does not change. A fully connected layer with an output of 224-dimensional geometric feature is added to the convolutional layer. We add L2 loss (weighted by 0.05) for the geometric features and the inferred features from the ground truth. (2) We branch out the 3rd inception-residual module at the early stage and feed its output to a series of convolutional layers with 1×1 kernels. The numbers of output channels are scaled twice by a factor of 1/2 until it reaches to 32 channels, followed by a fully connected layer. We add an L2 loss term (weighted by 0.05) for the injected edge features.

We denote the pixel location of the j-th body joint, $Y_j \in \mathcal{Z} \subset \mathbb{R}^2$, where $\mathcal{P}$ is the set of all (u, v) pixel locations in the image coordinate system. Our goal is to predict the image locations $Y = (Y_1, \ldots Y_J)$ for all J joints. As illustrated in Fig. 8, the output heatmaps are of size $J \times 64 \times 64$, denoted by $H = (H_1, \ldots, H_J)$, which are predicted beliefs for assigning a pixel location to each joint $Y_j = p, \forall z \in \mathcal{P}$, producing belief scores S_j for all pixels in the heatmap of joint j:

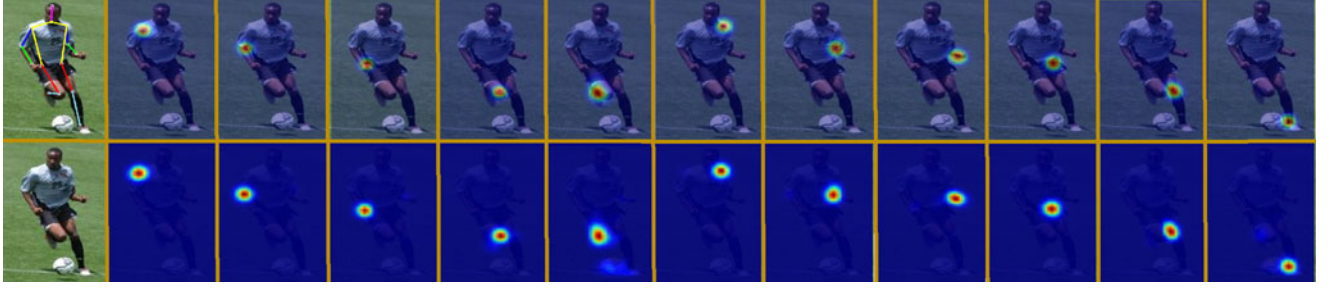$$H_j(p) \leftarrow S_j(Y_j = p) \qquad (6)$$

Fig. 8. Example output produced by our network. On the top-left, we see the final pose estimate provided by NMS across all heatmaps. Elsewhere we show sample heatmaps: (1) The first row shows the final part regression heatmap results; (2) the second row shows the preliminary part regression results from the intermediate supervision layer. The heatmaps from the first row have finer predictions than the second row, especially the heatmap for the right foot, where the preliminary prediction renders belief scores for the soccer ball as well.

In our experiments, we regress the input RGB images into a set of 15 heatmaps, 14 of which are human joints while the other one as the background. The heatmaps are then suppressed into joint locations Y using our proposed 3D-NMS algorithm specially designed for human pose estimation, which will be explained later. During training, we provide ground truth heatmaps for each joint by creating Gaussian peaks at ground truth locations. The cost function $\mathcal{L}_f$ we aim to minimize for the fractal network is given by:

$$\mathcal{L}_f = \sum_{j \in J} ||H_j(p) - H_j^*(p)||_2^2 \qquad (7)$$

The overall loss for training is a weighted combination of heatmap cost and projection matrix fitness provided by knowledge-guided learning, with a control parameter on how much guidance should be imposed. The overall network is then trained to minimize the following joint loss function:

$$W_f^* \leftarrow \arg\min_{W_f}(\lambda \times \mathcal{L}_{KP} + (1 - \lambda) \times \mathcal{L}_f) \qquad (8)$$

where $\mathcal{L}_{KP}$ and $\mathcal{L}_f$ are loss from knowledge projection layer and the fractal network loss, λ is the weight parameter decaying during training, and W_f^* is the trained parameters in the fractal network. The output of knowledge projection layer will guide the training of fractal network by generating a strong and explicit gradient applied to backward path to the injection layer in the following form:

$$\Delta W_{f,i} = -\lambda \cdot \frac{\partial \mathcal{L}_{KP}}{\partial W_{f,i}} \qquad (9)$$

Where $W_{f,i}$ is the weight matrix of injection layer in fractal network. Note that the network update only occurs during training. During testing, the knowledge representation and projection modules are removed.

E. Cross-Heatmap Nonmaximum Suppression

In this work, we introduce a novel pose non-maximum suppression (NMS) algorithm specially designed for human pose estimation. Our experiments in Section V-D show that employing pose-NMS consistently render better predictions for all models across iterations on both the MPII [60] and LSP [61] datasets. Instead of finding the maximum value at the pixel-level to predict joint location as in [18], [50], [51], [54], we detect blobs with strong responses in each heatmap. Basically, we gather

---

Algorithm 1: Summary of Procedures: Training Phase

> **input** : A set of RGB images I and corresponding ground truth joint coordinates J
> **output:** Trained weights W_f^* for the Fractal Network, W_{KP}^* for the knowledge projection layers
> **1** Initialize DNN with fractal network and knowledge projection layers;
> **2 for** k *epoches* **do**
> **3**   **for** *mini-batch in* I **do**
> **4**   Compute external knowledge representation: $K \leftarrow \{I_n, J_n\}$;
> **5**   Back-propagate w.r.t W_f, W_{KP} ;
> **6** $\{W_f', W_{KP}'\} \leftarrow$ $\arg\min_{W_f}(\lambda \times \mathcal{L}_{KP} + (1 - \lambda) \times \mathcal{L}_f)$;
> **7**   **end**
> **8 end**
> **9 return** $\{W_f^*, W_{KP}^*\}$

---

blobs from all heatmaps for suppression. We first find the blob with maximum response, then suppress other blobs from the same heatmap, and blobs from other heatmaps very close to this blob in the image coordinate system. We repeat this procedure until all blobs are removed. The suppression operates in the image coordinate system and channel-wise (u, v, c), therefore called cross-heatmap NMS.

IV. SUMMARY OF TRAINING AND TESTING PROCEDURES

We summarize our training and testing procedures in Algorithm 1 and 2, respectively. There are about 250 convolutional layers in the original hourglass network, while the proposed network with inception-resnet modules consist of over 300 convolutional layers. The network for training the proposed network has an additional cost with one external feature extraction module, two fully connected layers, three convolutional layers and two additional loss layers. In our implementation, the hourglass network takes an average of 47 ms to feed forward one a single Pascal TitanX GPU. The feed forward time of the proposed network with inception-resnet modules during testing is about 62 ms.

Algorithm 2: Summary of Procedures: Testing Phase

> **input** : A set of RGB images I and a fractal network with trained weights W_f^*
>
> **output:** A set of predicted joint coordinates J in the same image coordinate system

1 initialize network only with fractal network layers W_f^* ;
2 **while** *not at end of this image set* **do**
3   Load image I_i ;
4   Forward the network: $J_i \leftarrow \{W_f^*, I_i\}$;
5 **end**
6 return J

TABLE I
COMPONENT ANALYSIS ON THE LSP DATASET OF PCK@0.2 SCORE

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total
Hourglass	97.0	93.0	88.8	85.6	92.2	93.0	90.9	**91.5**
Ours (no guidance)	97.9	93.2	89.1	86.4	94.5	93.8	92.9	**92.6**
Ours (with guidance)	98.2	94.4	91.8	89.3	94.7	95.0	93.5	**93.9**
Plain testing	97.4	92.7	88.8	86.7	92.2	93.8	92.2	92.0
+ flipping	97.7	93.3	90.4	87.5	93.2	94.2	92.8	92.7
+ scaling	98.1	93.7	91.3	88.7	94.0	94.6	93.2	93.4
+ 3D-NMS	98.2	94.4	91.8	89.3	94.7	95.0	93.5	**93.9**

Note that numbers in bold indicate the method has employed all techniques during testing.

V. EXPERIMENTAL RESULTS

For comprehensive experimental analysis, we will first introduce the datasets, evaluation criteria and implementation details. Then we will present quantitative evaluations on benchmark datasets. Finally, diagnostic experiments, algorithm performance analysis and dicussions are provided for further analysis.

A. Datasets and Criteria

1) Datasets: We evaluate the proposed method on two widely used benchmark datasets: the MPII Human Pose [60] and extended Leeds Sports Poses (LSP) [61]. The MPII Human Pose dataset includes about 25K images with 40K annotated poses. The images are collected from YouTube videos covering daily human activities with highly articulated human poses. The LSP dataset consists of 11K training images and 1K testing images of sports activities.

2) Criteria: Three widely used criteria are used in our experiments to evaluate the performance of the proposed human pose estimation method: Percentage of Corrected Parts (PCP) [33], [62], [63], Percentage of Detected Joints (PDJ) [16], [19], [33], and Percentage of Corrected Keypoints (PCK) [33].

a) PCP: The PCP evaluates the localization accuracy of body parts (sticks of skeleton). It requires the estimated part end points must be within half of the part length from the ground truth part end points. As pointed by Yang and Ramanan [33], some previous work requires only the average of theendpoints of a part to be correct (PCP-average), rather than both endpoints (PCP-strict). Moreover, the early PCP implementation [62] selects the best matched output without penalizing false positives. In all our experiments, we adopt the strictest measure, i.e., PCP-strict with one single output, if not specially specified. For more detailed descriptions on PCP, please refer to [62] and [33].

b) PDJ: Though PCP is the initially preferred criterion for evaluation, it has the drawback of penalizing shorter limbs, such as lower arms. Thus the PDJ measure is introduced [16], [19] to measure the detection rate of body joints, where a joint is considered to be detected if the distance between the detected joint and the true joint is less than a fraction of the torso diameter. The torso diameter is usually defined as the distance between opposing joints on the human torso, such as left shoulder and right hip [16]. The Area Under Curve (AUC) can be used as the overall evaluation of the PDJ curve. In the following experiments, we report AUC as our PDJ performance.

c) PCK: The PCK measure is very similar to the PDJ criterion. The only difference is that the torso diameter is replaced with the maximum side length of the external rectangle of ground truth body joints. For full body images with extreme pose (especially when the torso becomes very small), the PCK may be more suitable to evaluate the accuracy of body part localization.

In our experiments, we follow the official benchmark evaluation protocals.[1] Official benchmark on MPII dataset adopts PCKh (using portion of head length as reference) at 0.5, while official benchmark on LSP dataset adopts both PCP and PCK at 0.2. LSP benchmark provide comparisons on both Observer-Centric (OC) and Person-Centric (PC) evaluations, of which the most widely adopted evaluation protocol is PCK-PC. In addition, both benchmarks adopt AUC scores.

B. Implementation Details

1) Data Augmentation: We crop the images with the target human centered at the images with roughly the same scale, and warp the image patch to the size 256×256. Then, we randomly rotate ($\pm 30°$) and flip the images, perform random re-scaling (0.75 to 1.25) and color jittering to make the model more robust to scale and illumination changes.

2) Experimental Settings: We use a modified version of *Caffe* [64] that produces three types of network outputs: the augmented image, the corresponding transformed ground truth heatmaps, and the injected knowledge for the augmented image. The knowledge projection is switched off during testing. We train our model using an initial learning rate of 2.5×10^{-4}. The parameters are optimized by the RMSprop [65] method. We divide the learning rate by 2 when the validation set hits plateaus. The minimum learning rate is set to 10^{-6}. We use 4 Pascal TITAN GPUs to train the model on the merged dataset of MPII and extended LSP for over 300 epochs, and adopt Tompson's validation split for the MPII dataset used in [17] to monitor the training process. The same model is used for the testing of both MPII and LSP test sets.

According to [66], we can use a prior condition for the background which forces the network to converge to zero. It is therefore important to scale the gradient responses using a weighting

---

[1] http://human-pose.mpi-inf.mpg.de/#evaluation

TABLE II
COMPARISONS OF PCKh@0.5 SCORE ON THE MPII TEST SET

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total
Ours	98.1	96.3	92.2	87.8	90.6	87.6	82.7	**91.2**
Newell *et al.*, ECCV'16 [18]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Bulat & Tzimiropoulos, ECCV'16 [54]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Wei *et al.*, CVPR'16 [51]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Insafutdinov *et al.*, ECCV'16 [46]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Rafi *et al.*, BMVC'16 [67]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Gkioxary *et al.*, ECCV'16 [68]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Lifshitz *et al.*, ECCV'16 [69]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Pishchulin *et al.*, CVPR'16 [45]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Hu & Ramanan, CVPR'16 [70]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Tompson *et al.*, CVPR'15 [44]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Carreira *et al.*, CVPR'16 [52]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson *et al.*, NIPS'14 [17]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Pishchulin *et al.*, ICCV'13 [34]	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1

TABLE III
COMPARISONS OF PCK@0.2 SCORE ON THE LSP TEST SET

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total
Ours	98.2	94.4	91.8	89.3	94.7	95.0	93.5	**93.9**
Bulat & Tzimiropoulos. ECCV'16 [54],	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Wei *et al.* CVPR'16 [51],	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Insafutdinov *et al.* ECCV'16 [46],	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Pishchulin *et al.* CVPR'16 [45],	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Lifshitz *et al.* ECCV'16 [69],	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Belagiannis & Zisserman FG'17 [50],	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2
Yu *et al.* ECCV'16 [71],	87.2	88.2	82.4	76.3	91.4	85.8	78.7	84.3
Rafi *et al.* BMVC'16 [67],	95.8	86.2	79.3	75.0	86.6	83.8	79.8	83.8
Yang *et al.* CVPR'16 [72],	90.6	78.1	73.8	68.8	74.8	69.9	58.9	73.6
Chen & Yuille NIPS'14 [43],	91.8	78.2	71.8	65.5	73.3	70.2	63.4	73.4
Fan *et al.* CVPR'15 [73],	92.4	75.2	65.3	64.0	76.7	68.3	70.4	73.0
Tompson *et al.* NIPS'14 [17],	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3
Pishchulin *et al.* ICCV'13 [34],	87.2	56.7	46.7	38.0	61.0	57.5	52.7	57.1
Wang & Li *et al.* CVPR'13 [74],	84.7	57.1	43.7	36.7	56.7	52.4	50.8	54.6

factor so that the contributions to the parameter updating process are equal between the foreground and background heatmap pixels. In our training process, we weight the foreground and background by 20 : 1. The neural network takes the cropped images patches or ROI of the images as inputs. However, the cropped patches or ROI might contain limbs from other persons. In this case, our ground truth simply ignores other limbs. For example, any region that is not from the keypoints of the target person is considered as background heatmap in the ground truth. Since the target person is always centered in the cropped image or ROI. Therefore, limbs from other persons usually have lower response in the predicted heatmaps.

3) Network Inference: During testing, we follow the standard routine to crop image patches with the given rough position and the scale of the test human for MPII dataset. For the LSP dataset, we use image size as the rough scale, and image center as the rough position of the target human to crop the image patches. Before feeding into the neural network, we further pre-process images with normalization and pixel-wise subtraction by estimated mean value. All the experimental results are produced from the original and flipped image pyramids with 2 scales (1 and 0.75). Note that we swap heatmaps
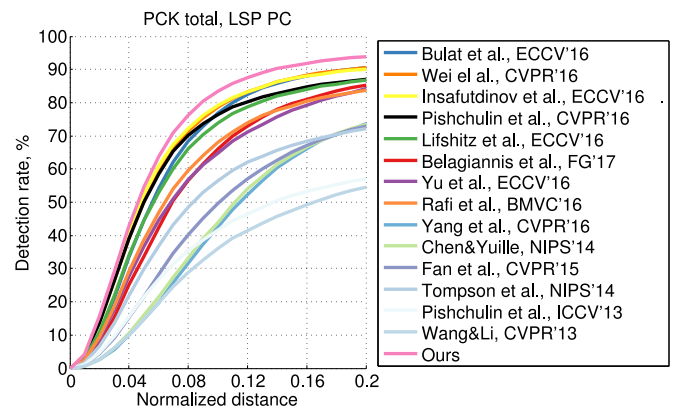

Fig. 9. Person-Centric (PC) PCK curves on the LSP test set. Ours is on top.

of left and right limbs before merging corresponding heatmaps for each joint. The merged heatmaps are transformed into joint coordinates by the proposed cross-heatmap non-maximum suppression method. The feed forward time of the network during testing is 62 ms with a single Pascal TITAN X GPU.

Fig. 10.   Qualitative results on the MPII test set.

C. Performance Evaluation

We use the Percentage Correct Keypoints (PCK) [33] metric for performance comparisons on the LSP dataset, and the PCKh measure [60], where the error tolerance is normalized with respect to the head size, for performance comparisons on the MPII Human Pose dataset. We train our model by adding the MPII training set to the extended LSP training set with person-centric (PC) annotations, which is a standard routine in the literature [45], [46], [50], [51].

1) Results on the MPII Human Pose Dataset: The AUC score of our network for MPII dataset is 63.6. Table II reports the comparison of the PCKh performance of our method with previous state-of-the-art at a normalized distance of 0.5. Our total PCKh-0.5 score achieves the state-of-the-art performance of 91.2%. Note that we test multiple scales (1 and 0.75) on both the MPII and LSP dataset. While cropping the images with the given scale of MPII dataset, for some images the feet are cropped out, therefore resulting in a lower detection rate for ankles. Qualitative results are illustrated in Fig. 10.

2) Results on the Leeds Sports Pose (LSP) Dataset: The AUC score of our network for LSP dataset is 69.1. Table III reports the PCK scores with a threshold of 0.2, and Fig. 9 shows the PCK scores with different thresholds. Our approach outperforms the state-of-the-art methods on each body part with a PCK score of 93.9%. Table IV reports the PCP scores at the threshold of 0.5. Qualitative results and remaining failure cases are illustrated in Figs. 11 and 12, respectively.

D. Algorithm Performance Analysis and Ablation Study

In the following, we perform component analysis of our proposed method on the LSP dataset. This analysis is not available for the MPII dataset since its ground truth data is not publicly available. We summarize the contribution of each component in Table I.

We compare the proposed inception-resnet module and the basic resnet module used by our stacked hourglass networks [18]. Since their performance is not specifically reported on the LSP dataset, we implement their network within our system to provide fair comparisons. With identical settings, our network with inception-resnet module achieves superior performance over that with basic resnet module, improving the accuracy by 1.1%. We also compare our network using standard training and knowledge-guided learning. Results show that better performance is achieved with knowledge-guided training with an accuracy improvement of 1.3%. We then analyze contributions of other techniques employed mainly during testing, i.e., flipping the images, testing the image at multiple scales, and using proposed NMS algorithm for pose estimation. Image flipping improves the performance by 0.7%, while testing on both original and 0.75 scales further improves performance by another 0.7%. Cross-heatmap non-maximum suppression improves the PCK score by 0.5%.

It should be noted that our implementation[2] in *PyCaffe* [64] may not fully reproduce equivalent performance on MPII dataset

[2]Code and models available at: http://github.com/Guanghan/GNet-pose

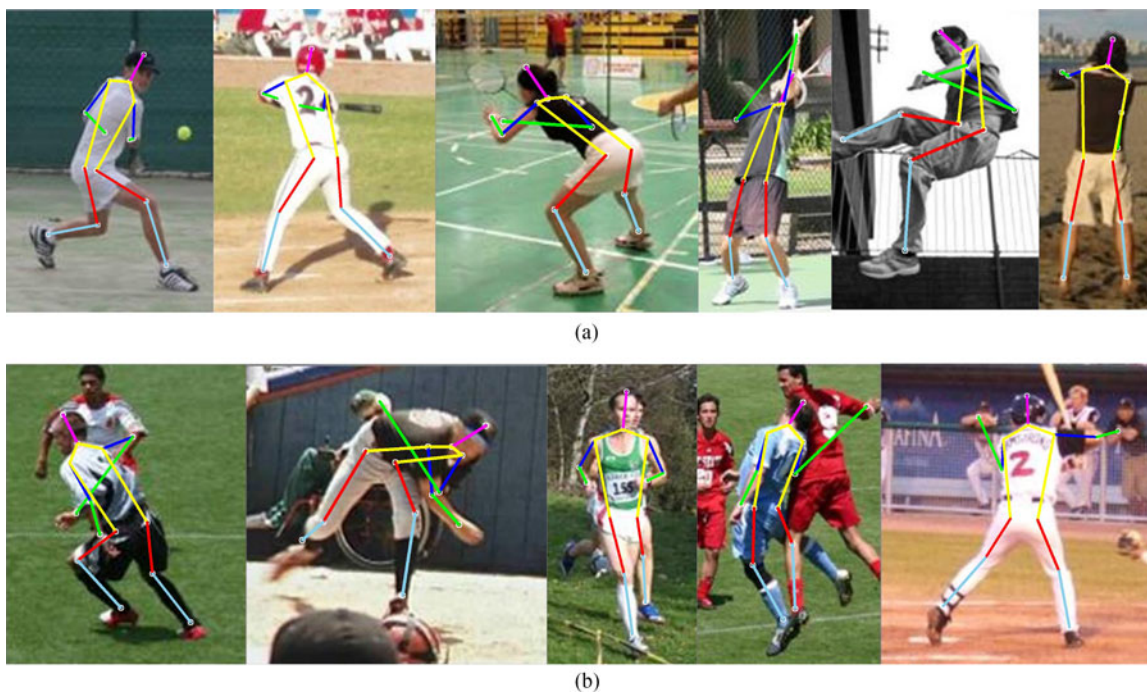Fig. 11. Qualitative results on the LSP test set.

(a)

(b)

Fig. 12. Failure cases on LSP dataset. (a) Ambiguity caused by full occlusion of two or more adjacent body parts. (b) Regression mistake caused by the concurrence of body part noise from other persons and full occulusion of less than two body parts.

TABLE IV
COMPARISONS OF PCP@0.5 SCORE ON THE LSP TEST SET

Method	Torso	U.Leg	L.Leg	U.Arm	Forearm	Head	Total
Ours	98.6	95.8	93.6	90.7	84.2	96.4	**92.3**
Bulat & Tzimiropoulos. ECCV'16 [54],	97.7	92.4	89.3	86.7	79.7	95.2	88.9
Wei *et al.* CVPR'16 [51],	98.0	82.2	89.1	85.8	77.9	95.0	88.3
Insafutdinov *et al.* ECCV'16 [46],	97.0	90.6	86.9	86.1	79.5	95.4	87.8
Yu *et al.* ECCV'16 [71],	98.0	93.1	88.1	82.9	72.6	83.0	85.4
Pishchulin *et al.* CVPR'16 [45],	97.0	88.8	82.0	82.4	71.8	95.8	84.3
Lifshitz *et al.* ECCV'16 [69],	97.3	88.8	84.4	80.6	71.4	94.8	84.3
Belagiannis & Zisserman FG'17 [50],	96.0	86.7	82.2	79.4	69.4	89.4	82.1
Rafi *et al.* BMVC'16 [67],	97.6	87.3	80.2	76.8	66.2	93.3	81.2
Yang *et al.* CVPR'16 [72],	95.6	78.5	71.8	72.2	61.8	83.9	74.8
Chen & Yuille NIPS'14 [43],	96.0	77.2	72.2	69.7	58.1	85.6	73.6
Fan *et al.* CVPR'15 [73],	95.4	77.7	69.8	62.8	49.1	86.6	70.1
Tompson *et al.* NIPS'14 [17],	90.3	70.4	61.1	63.0	51.2	83.7	66.6
Pishchulin *et al.* ICCV'13 [34],	88.7	63.6	58.4	46.0	35.2	85.1	58.0
Wang & Li *et al.* CVPR'13 [74],	87.5	56.0	55.8	43.1	32.1	79.1	54.1

of the hourglass network [18], which is implemented in *Torch* [75]. However, from our analysis, we can see that our proposed knowledge-guided training is able to improve the performance on top of existing deep neural network. We expect that the same performance gain can be achieved on other network structures.

VI. CONCLUSION

In this work, we proposed to encode and inject external human knowledge into deep neural networks to guide its training process with learned projections for more effective human pose estimation. We adopt the stacked hourglass design and propose to use inception-resnet as the building block of our fractal network to regress human pose into heatmaps with no explicit graphical modeling. Utilizing a multi-resolution feature representation with guided learning, the network learns an empirical set of low and high-level features which are typically more tolerant to variations in the training set. Knowledge-guided learning is a generic scheme that can be potentially used to aid other deep neural network training tasks. The effectiveness of the proposed inception-resnet module and the benefit in guided learning with knowledge projection is evaluated on two widely used human pose estimation benchmarks.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their insightful comments which helped improving the paper.

REFERENCES

[1] L. Fu, J. Zhang, and K. Huang, "ORGM: Occlusion relational graphical model for human pose estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 927–941, Feb. 2017.

[2] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Body parts dependent joint regressors for human pose estimation in still images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2131–2143, Nov. 2014.

[3] X. Zhang, C. Li, W. Hu, X. Tong, S. Maybank, and Y. Zhang, "Human pose estimation and tracking via parsing a tree structure based human model," *IEEE Trans. Syst., Man, Cybern.:* Syst., vol. 44, no. 5, pp. 580–592, May 2014.

[4] H. Jiang, "Human pose estimation using consistent max covering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1911–1918, Sep. 2011.

[5] L. Zhao, X. Gao, D. Tao, and X. Li, "Learning a tracking and estimation integrated graphical model for human pose tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 12, pp. 3176–3186, Dec. 2015.

[6] Q. Li, F. He, T. Wang, L. Zhou, and S. Xi, "Human pose estimation by exploiting spatial and temporal constraints in body-part configurations," *IEEE Access*, vol. 5, pp. 443–454, 2017.

[7] M. Eichner and V. Ferrari, "Human pose co-estimation and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2282–2288, Nov. 2012.

[8] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3D pictorial structures revisited: Multiple human pose estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 1929–1942, Oct. 2016.

[9] F. Zhou and F. De la Torre, "Spatio-temporal matching for human pose estimation in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1492–1504, Aug. 2016.

[10] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1913–1921.

[11] N. Ikizler-Cinbis and S. Sclaroff, "Web-based classifiers for human action recognition," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1031–1045, Aug. 2012.

[12] A. Marcos-Ramiro, D. Pizarro, M. Marron-Romera, and D. Gatica-Perez, "Let your body speak: Communicative cue extraction on natural interaction using RGBD data," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1721–1732, Oct. 2015.

[13] X. Cai, W. Zhou, L. Wu, J. Luo, and H. Li, "Effective active skeleton representation for low latency human action recognition," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 141–154, Feb. 2016.

[14] R. Ren and J. Collomosse, "Visual sentences for pose retrieval over low-resolution cross-media dance collections," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1652–1661, Dec. 2012.

[15] H. Kadu and C.-C. J. Kuo, "Automatic human mocap data classification," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2191–2202, Dec. 2014.

[16] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 1653–1660.

[17] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[18] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.

[19] B. Sapp and B. Taskar, "MODEC: Multimodal decomposable models for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2013, pp. 3674–3681.

[20] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2013, pp. 588–595.

[21] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 723–730.

[22] Y. Tian, C. L. Zitnick, and S. G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 256–269.

[23] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2013, pp. 3041–3048.

[24] L. Karlinsky and S. Ullman, "Using linking features in learning non-parametric part models," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 326–339.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 770–778.

[27] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 1–9.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 2818–2826.

[29] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2654–2662.

[30] A. Romero *et al.*, "FitNets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[31] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.

[32] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Trans. Comput.*, vol. C-22, no. 1, pp. 67–92, Jan. 1973.

[33] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[34] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 3487–3494.

[35] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 33–47.

[36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2005, vol. 1, pp. 886–893.

[37] X. Zhu, C. Vondrick, D. Ramanan, and C. C. Fowlkes, "Do we need more training data or better models for object detection?" in *Proc. Brit. Mach. Vis. Conf.*, 2012, vol. 3, p. 5.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[39] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 779–788.

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

[41] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[42] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2830–2838.

[43] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 1736–1744.

[44] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015 pp. 648–656.

[45] L. Pishchulin *et al.*, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 4929–4937.

[46] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 34–50.

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 3431–3440.

[48] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2018–2025.

[49] L. Wang, C.-Y. Lee, Z. Tu, and S. Lazebnik, "Training deeper convolutional networks with deep supervision," *arXiv preprint arXiv:1505.02496*, 2015.

[50] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," *arXiv preprint arXiv:1605.02914*, 2016.

[51] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 4724–4732.

[52] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 4733–4742.

[53] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2015, vol. 2, p. 5.

[54] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 717–732.

[55] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The difficulty of training deep architectures and the effect of unsupervised pre-training." in *Proc. Int. Conf. Artif. Intell. Stat.*, 2009, vol. 5, pp. 153–160.

[56] C. Bucilu, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 535–541.

[57] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[58] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 1685–1692.

[59] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.

[60] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 3686–3693.

[61] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 12.1–12.11.

[62] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2008, pp. 1–8.

[63] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," *Int. J. Comput. Vis.*, vol. 99, no. 2, pp. 190–214, 2012.

[64] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[65] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[66] P. H. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 82–90.

[67] U. Rafi, I. Kostrikov, J. Gall, and B. Leibe, "An efficient convolutional network for human pose estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 109.1–109.11.

[68] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 728–743.

[69] I. Lifshitz, E. Fetaya, and S. Ullman, "Human pose estimation using deep consensus voting," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 246–260.

[70] P. Hu and D. Ramanan, "Bottom-up and top-down reasoning with hierarchical rectified Gaussians," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 5600–5609.

[71] X. Yu, F. Zhou, and M. Chandraker, "Deep deformation network for object landmark localization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 52–70.

[72] W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 3073–3082.

[73] X. Fan, K. Zheng, Y. Lin, and S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 1347–1355.

[74] F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2013, pp. 596–603.

[75] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A MATLAB-like environment for machine learning," *BigLearn NIPS Workshop*, No. EPFL-CONF-192376, 2011.

Guanghan Ning received the B.S. degree in electronic and information engineering from Beijing Jiaotong University, Beijing, China, in 2012, and the M.S degree in electrical and computer engineering in 2013 from the University of Missouri, Columbia, MO, USA, where he is currently working toward the Ph.D. degree in electrical and computer engineering. He has developed various methods and software tools for license plate recognition, text detection from natural scenes, camera take detection, and content-based multimedia retrieval. His current research interests include human pose estimation, object detection, and tracking.

Zhiquan He received the Ph.D. degree in computer science from the University of Missouri, Columbia, MO, USA, in 2014. He is currently an Assistant Professor with the College of Information Engineering, Shenzhen University, Shenzhen, China. He was a Research and Development Software Engineer with MathWorks USA from 2014 to 2017 before joining Shenzhen University. His research interests include computer vision, machine learning, and deep learning.

Zhi Zhang received the B.S. degree in electronic and information engineering from Beijing Jiaotong University, Beijing, China, in 2012, and the M.S. degree in electrical and computer engineering in 2014 from the University of Missouri, Columbia, MO, USA, where he is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering. His current research interests include multimedia content retrieval, object recognition and classification, wild animal detection, and person reidentification.