# Human Pose Estimation with Parsing Induced Learner

Xuecheng Nie[1] Jiashi Feng[1] Yiming Zuo[2] Shuicheng Yan[1,3]
[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Department of Electronic Engineering, Tsinghua University, Beijing, China
[3]Qihoo 360 AI Institute, Beijing, China
niexuecheng@u.nus.edu elefjia@nus.edu.sg zuoym15@mail.tsinghua.edu.cn yanshuicheng@360.cn

## Abstract

*Human pose estimation still faces various difficulties in challenging scenarios. Human parsing, as a closely related task, can provide valuable cues for better pose estimation, which however has not been fully exploited. In this paper, we propose a novel Parsing Induced Learner to exploit parsing information to effectively assist pose estimation by learning to fast adapt the base pose estimation model. The proposed Parsing Induced Learner is composed of a parsing encoder and a pose model parameter adapter, which together learn to predict dynamic parameters of the pose model to extract complementary useful features for more accurate pose estimation. Comprehensive experiments on benchmarks LIP and extended PASCAL-Person-Part show that the proposed Parsing Induced Learner can improve performance of both single- and multi-person pose estimation to new state-of-the-art. Cross-dataset experiments also show that the proposed Parsing Induced Learner from LIP dataset can accelerate learning of a human pose estimation model on MPII benchmark in addition to achieving outperforming performance.*

## 1. Introduction

Human pose estimation is a fundamental task in computer vision, aiming to estimate joint locations of human body. Recent years have witnessed many efforts made to push its performance frontier [11, 25, 36, 40], but it remains challenging to apply it to realistic scenarios. Distracting factors, *e.g.* occlusion, self-similarity, large deformation, huge variation in pose configuration and appearance, often lead to inaccurate joint localization and even false joint categorization. As shown in Figure 1 (a), partial occlusion on left arm causes inaccurate localizations of left wrist, while high similarity between left and right legs results in false categorization of right ankle.

Human body parts, generated by human parsing methods [21, 23, 37], can provide useful contextual cues to help localize body joints in these challenging scenarios. For instance, when body part cues (from left-lower arm and right-lower leg, as shown in Figure 1 (b)) are taken into account in joint localization, estimation errors on the joints of left wrist and right ankle can be effectively corrected, as shown in Figure 1 (c). Motivated by this, some research works [13, 20, 38, 39] exploit the parsing information to help improve pose estimation performance. However, they generally perform human body parsing and pose estimation *separately* and utilize parsing results to refine body joint localization as post processing. Although some improvement has been achieved, they do not fully utilize parsing information in an effective and efficient way, thus suffer several limitations. First, they hand-craft features from parsing results which are not powerful or robust to large pose variations in the wild. Second, they only use parsing information for inference other than learning pose models and therefore do not strengthen pose models essentially. Third, their overall frameworks are not end-to-end learnable.

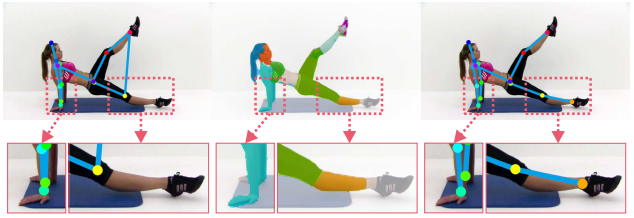



Figure 1. Illustration of our motivation for the proposed Parsing Induced Learner. (a) Pose estimation result without exploiting parsing information. (b) Parsing information generated from the proposed PIL. (c) Pose estimation result with the proposed PIL. The proposed PIL effectively leverages parsing information to refine the inaccurate locations and correct false categorizations for the highlighted body joints.

In this work, we propose to leverage human parsing information more effectively and efficiently for learning better pose estimation models and improving their performance. Targeting at the above limitations of existing works, we make following observations. First, the parsing representations should be learned towards being beneficial to pose estimation, instead of solely learned from the parsing supervision.

Second, the learned parsing representations should be effectively transferable to the pose estimation domain. Traditional multi-task learning frameworks [3, 12] keep the architecture for two tasks tied, which blurs the feature distinctiveness for pose estimation and parsing and limits their mutual benefits. Third, the pose estimation model should be dynamic and can fast adapt to various testing samples of different characteristics, relying on the transferred parsing information. Furthermore, human parsing annotation is available [16, 38] nowadays, providing us easy access to such information.

According to the above observations, we design a novel Parsing Induced Learner (PIL) that learns to fast adapt the pose estimation model conditioned on the parsing information extracted from a specific sample, and therefore effectively improves both performance and flexibility of the model. The proposed PIL refines inaccurate localization and corrects false categorization of body joints effectively, which are difficult to address for a static pose estimation model. As shown in Figure 1, PIL can exploit body part cues to constrain joint location and pose structure.

In particular, PIL consists of two components: an encoder that encodes an input image into high-level parsing representations, and an adapter that learns to adapt parameters of the pose model by leveraging parsing representations. The adaptive parameters predicted by PIL can help the pose model learn more tailored representations for estimating poses for each specific input, in which body part cues are effectively integrated for constraining joint locations and pose structures. Moreover, PIL can efficiently learn to adapt pose parameters in one-shot manner, yielding fast adaption of the base pose model according to parsing information. We implement PIL by combining a parsing encoder network and a parameter adapter network, which can be directly applied to various deep pose models and across different datasets. The parameter adapter network is trained with supervision from human pose to learn adaptive parameters for boosting pose estimation. The parsing encoder is learned from both the human parsing and pose annotations. Our whole model integrating pose estimation and PIL is end-to-end learnable.

Comprehensive experiments on popular benchmarks show that the PIL effectively improves performance for both single- and multi-person pose estimation to new state-of-the-art. We also conduct cross-dataset experiments to demonstrate its generalizability and transferability of exploiting parsing information from one dataset to another. Our contributions are three-fold. Firstly and most importantly, we propose a novel Parsing Induced Learner for efficiently learning to adapt pose estimation models by exploiting parsing information, achieving better pose estimation performance. Secondly, the proposed PIL is transferable across datasets, verified via experiments to apply the PIL trained on LIP dataset to MPII dataset, achieving both performance improvement and learning acceleration. Thirdly, with the help of PIL, an Hourglass network [25] based pose estimation model achieves new state-of-the-art on multiple benchmarks for both single- and multi-person pose estimation.

## 2. Related Work

Recently, research efforts have been devoted to both single- and multi-person pose estimation problems, via exploring network architecture engineering [25, 36, 40], enhancing training supervision [9, 10], and improving inference strategy [4, 14, 26]. However, they are still challenged by some distracting factors, *e.g.*, occlusion and self-similarity causing inaccurate joint locations and false joint categorization. Human body part cues from human parsing methods [7, 21, 23, 24, 37] can provide useful guidance to address the above challenges for constraining joint locations and pose structures. Motivated by this, some works [13, 20, 38, 39] have exploited parsing information to improve the performance of pose estimation.

In [39], Yamaguchi *et al.* proposed to use the normalized histograms of parsing labels around each location as additional features for refining the pose estimation results. In [20], Ladicky *et al.* proposed to use semantic segmentation of body parts to provide information on the appearance and shape of body joints. In [13], Dong *et al.* proposed the Grid Layout Feature to model the pairwise geometry relations between semantic parts and mixtures of joint-group templates, and then constructed the "And-Or" graph for simultaneously estimating joint locations and semantic labels. In [38], Xia *et al.* proposed to utilize semantic segmentation results to formulate additional features as the segment-joint smoothness term to encourage semantic and spatial consistency between parts and joints, and they modeled the multi-person pose estimation problem as a fully-connected conditional random field and solved it based on an Integer Linear Programming.

Despite the success of these existing works, they suffer from some obvious drawbacks on feature extraction from parsing information and guidance exploitation to learn pose models, which hamper them from sufficiently leveraging body part cues to estimate joint allocations. In contrast, our proposed Parsing Induced Learner can utilize parsing information to directly learn to extract features beneficial to pose estimation models rather than hand-crafting them. In addition, it can be integrated into the learning procedure of human pose estimation models by auxiliary parsing supervision. Moreover, our whole framework can be efficiently end-to-end learnable.

## 3. The Proposed Approach

### 3.1. The Formulation

Given an input RGB image $I \in \mathbb{R}^{M \times N \times 3}$ of size $M \times N$, our goal is to detect the locations $P = \{(x_i, y_i)\}_{i=1}^{J}$ of human

body joints with assistance of the corresponding (estimated) human parsing map $S \in \{0, 1, \ldots, L\}^{M \times N}$ of $I$. Here, $(x_i, y_i)$ are coordinates of the $i$th joint, and $J$ and $L$ are the number of joint and body part categories, respectively. In particular, 0 in $S$ denotes the background category.

Existing works [13, 20, 38, 39] suggest applying the parsing map $S$ in post processing to refine the pose estimation $P$. However, such a strategy does not essentially refine or enhance the pose estimation model. Differently, we propose a generic parsing induced pose estimation model $f_{[\theta,\theta']}$ (parameterized by $\theta$ and $\theta'$ together) to fully leverage parsing information learned from the pair $(I, S)$ in a flexible and effective way for getting more accurate pose estimation $P$, which is formulated as

$$f_{[\theta,\theta']} : I \rightarrow P, \text{ where } \theta' = g(I, S). \tag{1}$$

The above formulation features the essential difference between our proposed pose estimation model and existing ones. We enforce a part of the pose model parameters $\theta'$ to explicitly depend on the parsing map $S$ for better leveraging the parsing information through the function $g(\cdot, \cdot)$. Different parsing maps $S$ will induce different parameters $\theta'$ and thus modify the pose model $f_{[\theta,\theta']}$ dynamically. In this way, the pose model can be adaptive to the input image $I$ and parsing result $S$ fast and favorably through learning a proper model parameter prediction $g(\cdot, \cdot)$ end-to-end.

In particular, inspired by the "learning to learn" framework [2], we design a Parsing Induced Learner (PIL) to learn a well-performing function $g(\cdot, \cdot)$ such that the predicted parameters $\theta'$ can tailor the pose model to each input image based on parsing information and provide better pose estimation results. The proposed PIL consists of a *parsing encoder* for extracting the parsing features and a *parameter adapter* for learning the dynamic parameters $\theta'$, denoted as $E^S_{\theta^S}(\cdot)$ and $K_\phi(\cdot)$, respectively.

In PIL, the parameter adapter $K_\phi(\cdot)$ takes in the features output by the parsing encoder $E^S_{\theta^S}(\cdot)$ and predicts proper parameters $\theta'$ for the pose estimation model. Namely,

$$\theta' = g(I, S) := K_\phi\left(E^S_{\theta^S}(I)\right).$$

The pose estimation model $f_{[\theta,\theta']}$ contains an adaptive pose encoder $E^P_{[\theta^P,\theta']}(\cdot)$, which adopts the predicted parameter $\theta'$ from the PIL model and the other global parameter $\theta^P$ to output good features for body joint localizations. Introducing such a PIL model provides dynamic parameters $\theta'$and enables the adaptive pose encoder to fully exploit parsing information to extract better features for more accurate pose estimation through efficiently adapting its model parameters to specific input in one-shot.

On top of these two encoders are pose and parsing classifiers $C^P_{w^P}(\cdot)$ and $C^S_{w^S}(\cdot)$ which output the final pose and parsing estimations. In particular, $w^P$ and $\theta^P$ together instantiate $\theta$ in Eqn. (1). Given pose and parsing annotations $\widehat{P}$ and $\widehat{S}$, to jointly learn PIL with the pose and parsing models, we define the following loss function for training:

$$\mathcal{L} := \mathcal{L}^P\left(C^P_{w^P}(E^P_{[\theta^P,\theta']}(I)), \widehat{P}\right) + \beta\mathcal{L}^S\left(C^S_{w^S}(E^S_{\theta^S}(I)), \widehat{S}\right), \tag{2}$$

where $\mathcal{L}^P$ and $\mathcal{L}^S$ represent the pose and parsing loss functions respectively, defined in Sec. 3.3, and $\beta$ is a trade-off coefficient. Below, we will explain the implementation details for each component in our proposed approach.

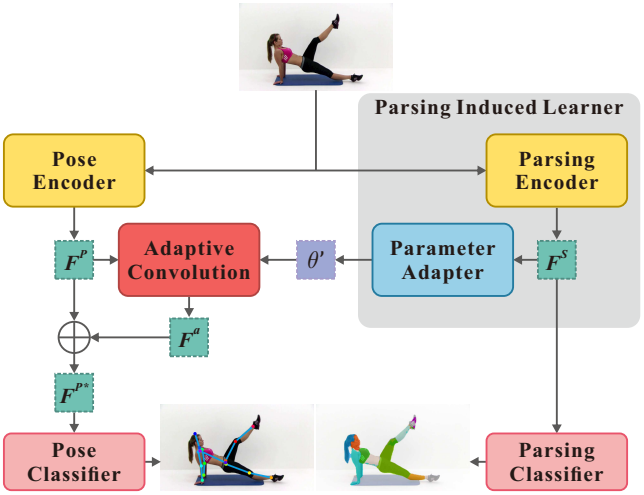



Figure 2. Overall architecture of our model. Given an input image, our model first utilizes a pose encoder to extract pose features $F^P$ and the proposed PIL to predict dynamic parameters $\theta'$ through a parameter adapter taking in parsing features $F^S$ from a parsing encoder. Then, our model feeds $F^P$ and $\theta'$ to an adaptive convolution to extract parsing induced features $F^a$ for fast adaption of the pose model. Our model regards $F^a$ as residual information and fuses it with $F^P$ via addition, leading to the refined features $F^{P*}$ for body joint localization. Finally, our model inputs $F^{P*}$ and $F^S$ to pose and parsing classifiers, respectively, to produce pose estimation and parsing prediction (ignored during testing).

## 3.2. The Network Architecture

Our implementation is based on deep Convolutional Neural Networks (CNNs). The overall architecture is shown in Figure 2. We now explain each component in details.

**Pose Encoder** The pose encoder $E^P_{\theta^P}(\cdot)$ extracts discriminative features $F^P = E^P_{\theta^P}(I)$ from the input image $I$ for pose estimation. We implement it via CNNs and evaluate two different network architectures in this work: one is the VGG16 based Fully Convolutional Network (FCN) [32] and the other is the state-of-the-art Hourglass network [25]. For the VGG16 based FCN architecture, we further remove the last two max pooling layers to reduce its total stride from 32 to 8 for more accurate joint localizations. For the Hourglass network, we follow the configurations in [25].

**Parsing Encoder** The parsing encoder $E^S_{\theta^S}(\cdot)$ is one component of the PIL model, aiming to extract features $F^S =$

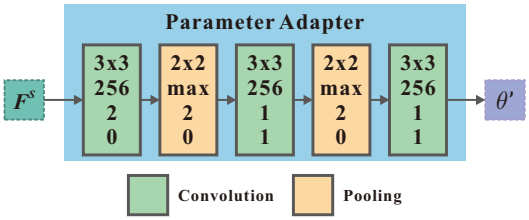



Figure 3. The architecture of the parameter adapter in PIL. The parameter adapter takes the features $F^S$ from the parsing encoder as input and outputs the adaptive convolution parameters $\theta'$. It is composed by stacking three convolution layers and two pooling layers. For each layer, the kernel size, the number of channels/pooling type, stride and padding size are specified from top to bottom.

$E^S_{\theta^S}(I)$ for encoding useful information for both parsing and pose estimation. In our implementation, the parsing encoder is also based on CNNs. Similarly, we also explore it with two different network architectures: the VGG16 based FCN and the Hourglass network. Note the parsing encoder need not be the same as the pose encoder as they are independent.

**Parameter Adapter** The parameter adapter $K_\phi(\cdot)$ is the other component of the PIL model, which is a one-shot learner to predict the dynamic parameters $\theta'$ via taking in the output $F^S$ of $I$ from the parsing encoder network. We implement it by a small CNN with learnable parameters $\phi$. In particular, the parameter adapter network predicts certain convolutional kernels of the pose encoder network. Its architecture is shown in Figure 3. The tensor $\theta' \in \mathbb{R}^{h\times h\times c}$ output from the last layer of the parameter adapter network is taken as the predicted dynamic convolutional kernels. Here $h$=7 is the convolution kernel size and $c$=$c_i \times c_o$ is the number of channels to learn for adaptive convolution with $c_i$ and $c_o$ as the number of input and output channels, respectively.

In practice, however, it is infeasible for the parameter adapter to directly predict all convolution parameters due to their large scale. For instance, given input feature maps with 256 channels and output feature maps with 256 channels, the number of convolution filters to be predicted by the parameter adapter is 256×256. The large scale parameters to predict would cause high space and time cost, and may result in overfitting [2]. To avoid these issues, we perform the following factorization [2] on the dynamic convolutional kernels $\theta'$ to reduce the number of free parameters:

$$\theta' = U * \tilde{\theta} *_c V, \tag{3}$$

where $*$ denotes the convolution operation, $*_c$ is the channel-wise convolution, $U$ and $V$ are auxiliary parameters to learn for the adaptive convolution and are explained in the next part. $\tilde{\theta}\in\mathbb{R}^{h\times h\times c_i}$ are the actual parameters to predict by the parameter adapter, with smaller size than original $\theta'$ by a magnitude. Eqn. (3) is analogous to SVD, where $\tilde{\theta}$ can be seen as coefficients over the parameter bases $U$ and $V$.

**Adaptive Convolution** To make best use of dynamic convolutional kernels $\theta'$ from the parameter adapter for directly extracting features to assist human pose estimation, we apply $\theta'$ on the highest-level features $F^P$ generated from the pose encoder, resulting in an adaptive convolution layer. The adaptive convolution layer is similar with the traditional convolution layer, just with the static convolution kernels replaced by the predicted dynamic convolution kernels $\theta'$:

$$F^a = \theta' * F^P = U * \tilde{\theta} *_c V * F^P,$$

where $F^a$ denotes the extracted features by dynamic parameters $\theta'$, $U\in\mathbb{R}^{1\times1\times c_i\times c_o}$ and $V\in\mathbb{R}^{1\times1\times c_i\times c_i}$ are auxiliary learnable parameters. In particular, we ignore the bias parameters in the adaptive convolution layer, due to the residual feature fusion strategy explained in the next part. Different from traditional CNN based features, $F^a$ is extracted in an efficient way by the dynamic parameters $\theta'$ based on parsing information for a given input image, rather than previous hand-crafted parsing based features for human pose estimation. Moreover, the parameter basis $\theta'$ of the adaptive convolution layer can be efficiently learned by the proposed PIL in one-shot, getting rid of iteratively updating weights based on large training datasets. In the implementation for the adaptive convolution layer, given $F^P$, we first use a 1×1 convolution on it with $V$, then conduct the dynamic convolution by group with $\tilde{\theta}$, and finally adopt another 1×1 convolution with $U$ to generate $F^a$.

**Feature Fusion** Different with features $F^P$ from the pose encoder network, $F^a$ is extracted based on parsing information and complementary to $F^P$ for human pose estimation. Hence, we adopt the residue learning idea [17] to regard $F^a$ as a residue component, and fuse it with the original features $F^P$ via addition:

$$F^{P*} = F^P + F^a,$$

where $F^{P*}$ is the final feature refined by parsing information for human pose estimation.

**Classifiers** After generating the final feature $F^{P*}$ for human pose estimation, we exploit a linear classifier $C^P_{w^P}(\cdot)$ on it to generate the predicted confidence maps for each kind of joints, by implementing a 1×1 convolution on $F^{P*}$. Similarly, we exploit another linear classifier $C^S_{w^S}(\cdot)$ on $F^S$ to generate the parsing prediction.

### 3.3. Training and Inference

As defined in Eqn. (2), we introduce two supervision to train the overall network model. We use Mean Square Error loss as $\mathcal{L}^P$ for training the pose model and PIL, and use Cross Entropy loss as $\mathcal{L}^S$ together with $\mathcal{L}^P$ to train the parsing model. The overall model is end-to-end trainable by gradient backpropagation.

In fact, the PIL can also be pretrained on one dataset and directly applied to assist pose estimation on new datasets. In other words, the PIL is able to transfer acquired parsing

information across different application datasets. We verify this property of PIL in the experiments.

During inference, the pose and parsing encoder networks take in the same image. The parameter adapter learns the parsing related convolution kernels in one feed-forward pass. We ignore the predicted parsing results and only take the output from the pose classifier for human pose estimation. For single-person pose estimation, we directly output the positions with maximum responses for each type of body joints. For multi-person pose estimation, we perform NMS to find joint candidates on the predicted confidence maps.

# 4. Experiments

## 4.1. Experimental Setup

**Datasets** We evaluate our proposed model on three most popular benchmarks for human pose estimation: Look into Person (LIP) [16], extended PASCAL-Person-Part [38], and MPII Human Pose Single-Person (MPII) [1], ranging from single-person to multi-person pose estimation and presenting various challenging scenarios.

The LIP dataset is a large-scale single-person dataset providing both human pose and parsing annotations, including locations for 16 body joints and annotations for 19 semantic body parts with one background category. In total, there are 50,462 images, which are split into three subsets: 30,462 for training, 10,000 for validation, and 10,000 for testing.

The extended PASCAL-Person-Part dataset presents multi-person images with both pose and parsing annotations for 14 body joints and 6 body parts. The total 3,533 images are split into 1,716 for training and 1,817 for testing.

The MPII dataset is another large-scale benchmark for single-person pose estimation. It contains 19,185 training and 7,247 testing images but only provides pose annotations for 16 body joints. On this dataset, we aim to evaluate the cross-dataset generalizability and transferability of the proposed PIL, *i.e.*, how well the model learns useful and transferable parsing information from one dataset (LIP) to assist pose estimation on another new dataset (MPII).

**Data Augmentation** For the LIP and extended PASCAL-Person-Part datasets, we crop training samples on original images based on the person center. We augment each training sample with rotation degrees in $[-40°, 40°]$, scaling factors in $[0.8, 1.5]$, translational offset $[-40\text{px}, 40\text{px}]$, and horizontally mirror. For MPII dataset, we augment each training sample with rotation degrees in $[-30°, 30°]$, scaling factors in $[0.7, 1.3]$, and horizontally mirror, but no translation augmentation. We resize and pad training samples to size $256\times256$ before inputting to CNNs. These augmentations are common and also used by previous works for both single- and multi-person pose estimation [4, 19, 25, 36].

**Implementation** For the LIP and extended PASCAL-Person-Part datasets, we train the overall network from scratch on their individual training samples. Since PASCAL-Person-Part images contain multiple persons, we need to associate joint candidates to corresponding person instances. In experiments, we follow the approach in [26], which learns to allocate joints simultaneously with the joint detection model. For MPII dataset, we directly use the parsing encoder and parameter adapter trained on the LIP dataset as PIL without further fine-tuning. We train the pose estimation network and auxiliary parameters in the adaptive convolution layer from scratch using the training samples from this dataset. We implement the proposed model using PyTorch [27]. We use RMSProp [33] as the optimizer. The learning rate is initially set as 0.0025, and decreased by multiplying 0.5 at the 150th, 170th and 200th epoch. We train all the models for 250 epochs in total. Testing is conducted on six-scale image pyramids with flipping. Our code will be made available.

**Metrics** The PCK [41] and Mean Average Precision (mAP) [29] are used for performance evaluation on the LIP and extended PASCAL-Person-Part datasets, respectively. We use the official PCKh metric [1] for performance evaluation on MPII dataset, following conventions.

## 4.2. Results on LIP Dataset

**Ablation Analysis** To evaluate our proposed model, we investigate two different backbone networks on the LIP validation set: the prevalent VGG16 network [32] successfully applied to various computer vision tasks [6, 29, 31], and the state-of-the-art Hourglass network [25] for human pose estimation. We first evaluate our proposed Parsing Induced Learner (PIL) based on VGG16 and compare it with various popular strategies (including feature fusion through adding, multiplying and concatenating) on exploiting parsing features for pose estimation, in order to demonstrate its efficacy. The results are summarized in Table 1, where VGG16-PIL(VGG16) denotes our proposed full model with VGG16 for both pose and parsing encoder networks, and VGG16-Add/Multi/Concat represent the models using other parsing utilization strategies. We also compare the adaptive ability of PIL with the traditional multi-task learning framework [31, 42] for joint human parsing and pose estimation, which is implemented based on VGG16 for representation learning with both pose and parsing supervision, denoted as VGG16-MTL. To disentangle effects of the residual module followed pose encoder on the estimation performance from network architecture engineering, we also evaluate another variant of our model by removing the proposed PIL and replacing the adaptive convolution layer with the traditional convolution layer, denoted as VGG16-Self.

From Table 1, one can observe that the proposed VGG16-PIL(VGG16) improves the baseline VGG16 by $8.5\%$ in terms of the average PCK, from $69.1\%$ to $75.0\%$. This result clearly shows our proposed model is effective at exploiting parsing information to learn powerful representations for

Table 1. VGG16 based ablation studies on LIP validation set. The model in parenthesis denotes the parsing network used in PIL.

| | Head | Sho. | Elb. | Wri. | Hip | Knee | Ank. | U.Body | PCK |
|---|---|---|---|---|---|---|---|---|---|
| VGG16 | 88.0 | 80.0 | 68.0 | 69.5 | 48.9 | 60.2 | 64.5 | 76.6 | 69.1 |
| VGG16-Add | 88.1 | 79.5 | 68.0 | 69.9 | 48.3 | 59.7 | 61.7 | 76.6 | 68.6 |
| VGG16-Multi | 87.3 | 75.0 | 60.2 | 65.1 | 42.7 | 51.9 | 58.4 | 72.2 | 63.7 |
| VGG16-Concat | 88.2 | 77.7 | 64.1 | 67.4 | 44.0 | 55.2 | 61.0 | 74.6 | 66.1 |
| VGG16-MTL | 87.4 | 76.5 | 61.9 | 66.1 | 46.0 | 53.8 | 59.8 | 73.3 | 65.3 |
| VGG16-Self | 88.0 | 81.1 | 70.0 | 69.5 | 50.3 | 61.2 | 63.5 | 77.4 | 69.8 |
| VGG16-PIL(VGG16) | **90.0** | **83.3** | **75.2** | **75.0** | **57.0** | **69.3** | **72.0** | **81.1** | **75.0** |

Table 2. Hourglass based ablation studies on LIP validation set. The model in parenthesis denotes the parsing network used in PIL.

| | Head | Sho. | Elb. | Wri. | Hip | Knee | Ank. | U.Body | PCK |
|---|---|---|---|---|---|---|---|---|---|
| HG-1s-1u: | 91.7 | 86.7 | 80.6 | 77.4 | 67.8 | 73.5 | 69.0 | 84.3 | 78.8 |
| HG-1s-1u-PIL(HG-1s-1u) | 92.6 | 89.0 | 84.2 | 81.3 | 70.8 | 78.4 | 75.5 | 86.9 | 82.2 |
| HG-2s-1u: | 91.8 | 88.3 | 82.4 | 79.1 | 70.6 | 75.9 | 73.7 | 85.6 | 80.8 |
| HG-2s-1u-PIL(HG-1s-2u) | 92.7 | 89.8 | 84.9 | 81.6 | 72.8 | 79.2 | 76.7 | 87.4 | 83.0 |
| HG-4s-2u | 93.2 | 90.7 | 86.7 | 83.2 | 72.6 | 81.7 | 80.8 | 88.6 | 84.5 |
| HG-4s-2u-PIL(HG-1s-2u) | 93.3 | 91.1 | 87.5 | 84.7 | 73.5 | 82.6 | 81.9 | 89.3 | 85.3 |
| HG-8s-1u | 93.4 | 91.2 | 87.3 | 84.4 | 73.3 | 81.8 | 80.8 | 89.2 | 84.9 |
| HG-8s-1u-PIL(HG-1s-2u) | 93.3 | 91.3 | 88.0 | 85.1 | 73.5 | 83.5 | 82.1 | 89.5 | 85.6 |

Table 3. Experiments on impacts of the parsing performance (measured by mIOU) on human pose estimation in our model. The model in parenthesis denotes the parsing network used in PIL.

| | mIOU | Head | Sho. | Elb. | Wri. | Hip | Knee | Ank. | U.Body | PCK |
|---|---|---|---|---|---|---|---|---|---|---|
| HG-4s-2u | - | 93.2 | 90.7 | 86.7 | 83.2 | 72.6 | 81.7 | 80.8 | 88.6 | 84.5 |
| +PIL(HG-1s-1u) | 41.1 | 93.3 | 91.0 | 87.0 | 84.3 | 73.7 | 82.2 | 80.7 | 89.0 | 85.0 |
| +PIL(HG-1s-2u) | 42.4 | 93.3 | 91.1 | 87.5 | 84.7 | 73.5 | 82.6 | 81.9 | 89.3 | 85.3 |
| +PIL(HG-1s-4u) | 43.3 | 93.3 | 91.4 | 87.7 | 84.8 | 72.9 | 83.2 | 82.3 | 89.4 | 85.4 |
| +PIL(HG-1s-8u) | 43.9 | 93.2 | 91.2 | 87.9 | 85.1 | 73.4 | 83.4 | 82.1 | 89.5 | 85.5 |

assisting human pose estimation.

Comparing VGG16-PIL(VGG16) with VGG16-Add/Multi/Concat baselines again demonstrates the improvement is not simply due to using parsing features. Although accessing and fusing parsing features, VGG16-Add/Multi/Concat even harm the pose estimation performance. This demonstrates naive feature fusion is not an effective way of utilizing parsing information as expected. Traditional multi-task learning framework VGG16-MTL suffers performance decline on human pose estimation, showing that directly introducing parsing supervision in training cannot effectively adapt parsing information to the pose estimation model. Comparing with VGG16-MTL, our PIL can effectively adapt parsing information to both learning and inference processes of human pose estimation. Adding residual module to the VGG16 backbone as VGG16-Self improves performance incrementally (from $69.1\%$ to $69.8\%$). This confirms that the proposed PIL extracts valuable information from parsing for human pose estimation rather than benefiting from the network architecture engineering.

We conduct similar ablation analysis on the proposed model using state-of-the-art architecture, the Hourglass network [25], for human pose estimation. The results are shown in Table 2, in which HG-$m$s-$n$u denotes the Hourglass network consisting of $m$ stacked Hourglass modules and each module with $n$ unit depth (32 layers). HG-$m$s-$n$u-PIL (HG-$m'$s-$n'$u) denotes the model with the proposed PIL. We make such choices to comprehensively evaluate the Hourglass network and its counterparts with our proposed PIL on the LIP dataset, aiming to analyze the effects of the depth and stages of Hourglass network on the performance of our proposed model for human pose estimation.

Table 4. Comparison with state-of-the-arts on LIP testing set.

| | Head | Sho. | Elb. | Wri. | Hip | Knee | Ank. | PCK |
|---|---|---|---|---|---|---|---|---|
| Hybrid Pose Machine | 71.7 | 87.1 | 82.3 | 78.2 | 69.2 | 77.0 | 73.5 | 77.2 |
| BUPTMM-POSE | 90.4 | 87.3 | 81.9 | 78.8 | 68.5 | 75.3 | 75.8 | 80.2 |
| Pyramid Stream Network | 91.1 | 88.4 | 82.2 | 79.4 | 70.1 | 80.8 | 81.2 | 82.1 |
| Chou *et al.* [10] | 94.9 | 93.1 | 89.1 | 86.5 | 75.7 | **85.5** | **85.7** | 87.4 |
| Our model | **94.9** | **93.1** | **89.9** | **87.6** | **75.9** | 84.9 | 84.4 | **87.5** |

From Table 2, one can observe that the proposed model always brings performance improvement over the backbone networks even though their performance is already very high. We can also observe that although HG-1s-1u-PIL (HG-1s-1u) and HG-2s-1u have similar numbers of parameters, HG-1s-1u-PIL (HG-1s-1u) achieves superior performance $82.2\%$ PCK, compared with HG-2s-1u that achieves $80.8\%$ PCK, which also shows the efficiency and effectiveness of the proposed model in exploiting parsing information for human pose estimation. In addition, we can find that PIL with a smaller parsing network can also improve the performance. More importantly, the proposed model gives new state-of-the-art of $85.6\%$ PCK on the LIP validation dataset.

We also conduct ablation experiments to study how different parsing networks (with different parsing qualities) affect human pose estimation. We fix the pose network as HG-4s-2u, and increase the depth of the parsing network from HG-1s-1u to HG-1s-8u to obtain increasingly better parsing performance measured by Mean Intersection over Union (mIOU) [16] from $41.1\%$ to $43.9\%$. The results in Table 3 reveal the trend that better parsing performance gives better pose estimation, reflecting the proposed PIL is good at exploiting parsing information.

**Comparison with State-of-the-arts** We compare the proposed model HG-8s-1u-PIL(HG-1s-2u) with state-of-the-arts on the LIP testing set. The results are shown in Table 4. In particular, the model proposed in [10] wins the CVPR 2017 LIP Human Pose Estimation Challenge. It uses a self adversarial training strategy for refining the pose estimation. The other two models, BUPTMM-POSE and Hybrid Pose Machines, combine the predictions of Hourglass networks and convoutional pose machines. Without sophisticated refinement or model ensemble, our proposed model outperforms all these well established baselines and achieves new state-of-the-art $87.5\%$ PCK. The proposed model is superior to [10] for most body joints, though the adversarial training is slightly better at refining joints of knee and ankle. By the PCK measurement, there is a tolerance between the prediction and the groundtruth, and the estimated pose structure

Table 5. Experiments on the extended PASCAL-Person Part dataset.

| | Head | Sho. | Elb. | Wri. | Hip | Knee | Ank. | mAP |
|---|---|---|---|---|---|---|---|---|
| Chen and Yuille [8] | 45.3 | 34.6 | 24.8 | 21.7 | 9.8 | 8.6 | 7.7 | 21.8 |
| Insafutdinov *et al.* [19] | 41.5 | 39.3 | 34.0 | 27.5 | 16.3 | 21.3 | 20.6 | 28.6 |
| Xia *et al.* [38] | 58.0 | 52.1 | 43.1 | 37.2 | 22.1 | **30.8** | **31.1** | 39.2 |
| Our baseline (w/o PIL) | 66.2 | 54.8 | 43.8 | 40.2 | 23.7 | 24.9 | 23.5 | 39.6 |
| Our model | **67.8** | **56.6** | **45.7** | **41.9** | **24.2** | 26.4 | 24.2 | **41.0** |

is more important than the absolute joint location. Hence, our approach is not significantly superior to [10] under the PCK measurement, since the adversarial training strategy can also appropriately constrain the predicted pose structure. Nevertheless, our proposed model can help generate more accurate joint locations. From experiments on MPII dataset in Sec. 4.4, we can observe that our proposed model significantly outperforms [10] under the AUC measurement.

**Qualitative Results** We visualize some qualitative results in Figure 5 (a) to better show the effectiveness of the proposed model in exploiting parsing information. From the results, one can observe the PIL corrects false detections on elbows caused by occlusion, benefiting from left and right arm part cues. In addition, PIL helps recover the missed detection on right lower arm due to large pose variations in the second image.

### 4.3. Results on PASCAL-Person-Part Dataset

As the extended PASCAL-Person-Part dataset involves multi-person pose estimation, we re-implement the model proposed in [26] with Hourglass network as the backbone. In particular, the pose network adopts Hourglass with 8 stacked modules HG-8s-1u and the parsing network of PIL is a much smaller one with only 1 Hourglass module HG-1s-2u.

Our performance and comparison with state-of-the-arts are shown in Table 5. The vanilla baseline model (without PIL) achieves 39.6% mAP. Introducing the PIL improves the performance to 41.0% mAP, offering a new state-of-the-art on this dataset. Moreover, our proposed model outperforms the best performing baseline [38] by a margin 2% mAP. These results also demonstrate the strong generalizability of our proposed model from single-person to multi-person pose estimation domain.

Figure 5 (c) shows qualitative results for multi-person pose estimation. The proposed PIL effectively refines the joint localizations by better exploiting the part cues. In particular, it successfully helps isolate joints from neighboring persons that are easy to confuse (see the first and second examples). Moreover, PIL corrects the false joint detections, as shown in the third and forth examples.

### 4.4. Results on MPII Dataset

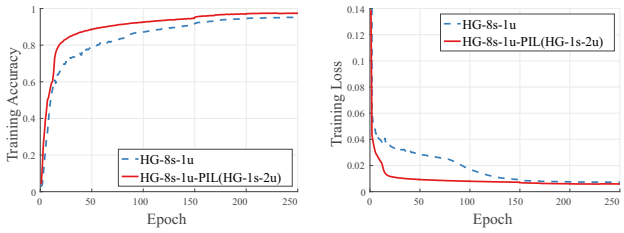
We consider a more challenging scenario where the PIL model is trained on a different dataset, aiming to evaluate the transferability of our proposed model on "learning to adapt" across different datasets. In the experiments, we use HG-8s-1u as the backbone pose network and HG-1s-2u as the parsing network of PIL. As MPII does not provide parsing annotations, we directly exploit the PIL trained on the LIP dataset. We fix the parameters of PIL to predict dynamic filters $\tilde{\theta}$ from LIP dataset, and learn auxiliary parameters $U$ and $V$ with the pose model, together.

Table 6. Ablation experiments on MPII validation set. The model in parenthesis denotes the parsing network used in PIL.

| | Head | Sho. | Elb. | Wri. | Hip | Knee | Ank. | U.Body | PCKh |
|---|---|---|---|---|---|---|---|---|---|
| HG-8s-1u | 97.7 | 96.2 | 90.6 | 86.3 | 89.8 | 85.9 | 82.1 | 92.7 | 90.2 |
| HG-8s-1u-PIL(HG-1s-2u) | **97.8** | **96.5** | **91.4** | **87.3** | **90.7** | **87.3** | **83.7** | **93.3** | **91.0** |

Table 7. Comparison with state-of-the-arts on MPII testing set.

| | Head | Sho. | Elb. | Wri. | Hip | Knee | Ank. | PCKh | AUC |
|---|---|---|---|---|---|---|---|---|---|
| Pishchulin *et al.* [28] | 74.3 | 49.0 | 40.8 | 34.1 | 36.5 | 34.4 | 35.2 | 44.1 | 24.5 |
| Tompson *et al.* [35] | 95.8 | 90.3 | 80.5 | 74.3 | 77.6 | 69.7 | 62.8 | 79.6 | 51.8 |
| Carreira *et al.* [5] | 95.7 | 91.7 | 81.7 | 72.4 | 82.8 | 73.2 | 66.4 | 81.3 | 49.1 |
| Tompson *et al.* [34] | 96.1 | 91.9 | 83.9 | 77.8 | 80.9 | 72.3 | 64.8 | 82.0 | 54.9 |
| Hu&Ramanan [18] | 95.0 | 91.6 | 83.0 | 76.6 | 81.9 | 74.5 | 69.5 | 82.4 | 51.1 |
| Pishchulin *et al.* [29] | 94.1 | 90.2 | 83.4 | 77.3 | 82.6 | 75.7 | 68.6 | 82.4 | 56.5 |
| Lifshitz *et al.* [22] | 97.8 | 93.3 | 85.7 | 80.4 | 85.3 | 76.6 | 70.2 | 85.0 | 56.8 |
| Gkioxary *et al.* [15] | 96.2 | 93.1 | 86.7 | 82.1 | 85.2 | 81.4 | 74.1 | 86.1 | 57.3 |
| Rafi *et al.* [30] | 97.2 | 93.9 | 86.4 | 81.3 | 86.8 | 80.6 | 73.4 | 86.3 | 57.3 |
| Insafutdinov *et al.* [19] | 96.8 | 95.2 | 89.3 | 84.4 | 88.4 | 83.4 | 78.0 | 88.5 | 60.8 |
| Wei *et al.* [36] | 97.8 | 95.0 | 88.7 | 84.0 | 88.4 | 82.8 | 79.4 | 88.5 | 61.4 |
| Newell *et al.* [25] | 98.2 | 96.3 | 91.2 | 87.1 | 90.1 | 87.4 | 83.6 | 90.9 | 62.9 |
| Chu *et al.* [11] | 98.5 | 96.3 | 91.9 | 88.1 | 90.6 | 88.0 | 85.0 | 91.5 | 63.8 |
| Chou *et al.* [10] | 98.2 | 96.8 | 92.2 | 88.0 | 91.3 | 89.1 | 84.9 | 91.8 | 63.9 |
| Chen *et al.* [9] | 98.1 | 96.5 | 92.5 | 88.5 | 90.2 | **89.6** | 86.0 | 91.9 | 61.6 |
| Yang *et al.* [40] | 98.5 | 96.7 | 92.5 | 88.7 | 91.1 | 88.6 | 86.0 | 92.0 | 64.2 |
| Our model | **98.6** | **96.9** | **93.0** | **89.1** | **91.7** | 89.0 | **86.2** | **92.4** | **65.9** |

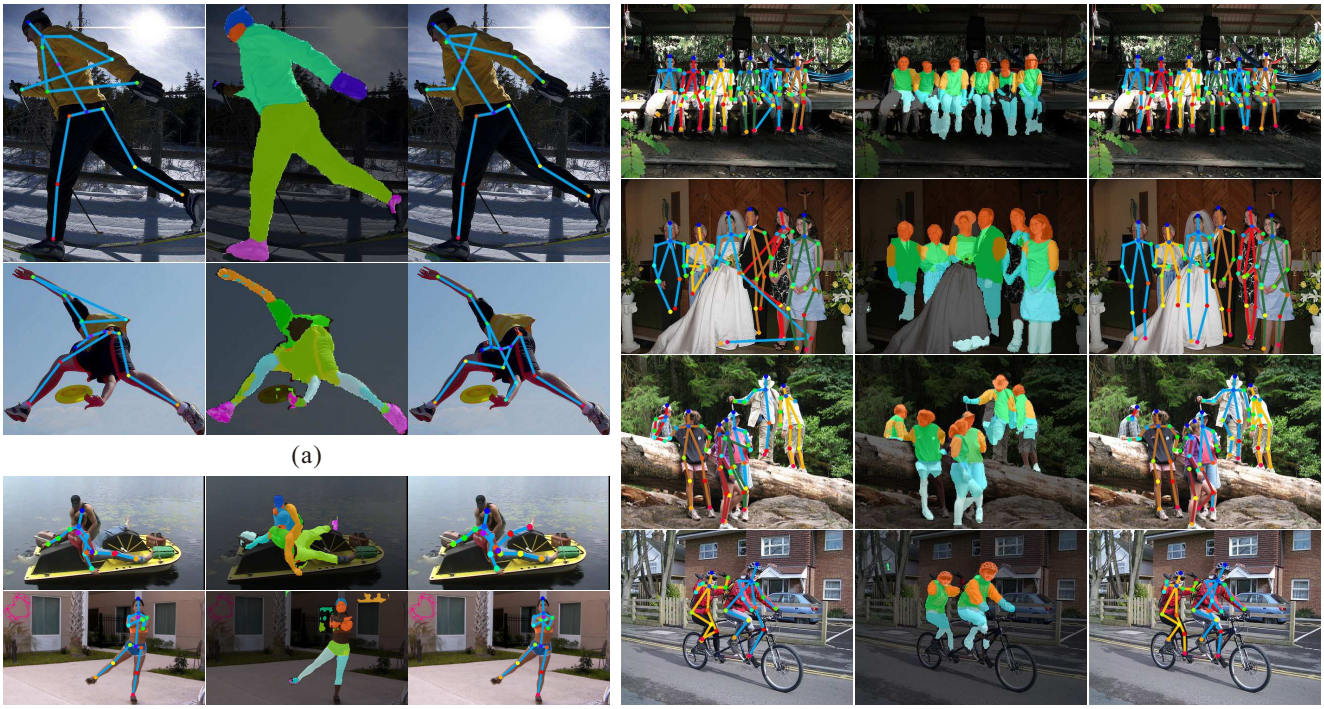



(a) (b)

Figure 4. Training accuracy and loss on MPII training set, shown in (a) and (b) respectively, to demonstrate the proposed PIL can accelerate learning speed of human pose estimation model.

**Ablation Analysis** We use the same validation set with Tompson *et al.* [34] to conduct the ablation analysis, and show results in Table 6. Our implementation of HG-8s-1u achieves 90.2% PCKh. Introducing PIL improves the performance to 91.0% PCKh. We also find that PIL improves prediction accuracy for all body joints, although it is trained on a different dataset. This demonstrates that our proposed model can successfully transfer useful parsing information from LIP dataset to MPII dataset.

In Figure 4, we also plot the training accuracy and loss of models with or without PIL on MPII training set. we can find



(a) (b) (c)

Figure 5. Qualitative results on (a) LIP dataset, (b) MPII dataset, and (c) extended PASCAL-Person-Part dataset. For each row, the left image shows the pose estimation result of the baseline model HG-8s-1u, the middle one shows the parsing map predicted by the proposed PIL, and the right one shows the pose estimation result of our proposed model HG-8s-1u-PIL(HG-1s-2u).

that both increase of training accuracy and decrease of loss go much faster with the help of PIL. These results demonstrate the effectiveness of the proposed PIL on accelerating the learning speed of human pose estimation model.

**Comparison with State-of-the-arts** Table 7 shows comparison of our model with state-of-the-arts. With the PIL from LIP dataset, our model achieves new state-of-the-art $92.4\%$ PCKh. One can also observe that the PIL improves the performance for most of the joints, except for the knee. The reason lies in the differences between LIP dataset and MPII dataset, which make the PIL model from LIP dataset unable to cover all variations in the configuration of knees in MPII dataset. Hence, our model cannot outperform [9, 10] with adversarial training for constraining joint configurations of human body. Moreover, our model significantly improves AUC over the best performing baseline from $64.2\%$ to $65.9\%$, showing PIL can indeed effectively utilize parsing information to better localize body joints.

**Qualitative Results** Qualitative results are shown in Figure 5 (b). One can observe that the PIL model trained on LIP dataset can perform well for some images on MPII dataset, *e.g.* the first example. In this case, PIL successfully transfers parsing information and corrects false detections on legs of the person due to self ambiguity. In the second example, the parsing model fails to generate high-quality parsing results. However, we surprisingly find that PIL is still able to provide useful cues to refine the pose estimation on hands. This shows the good generalizability of PIL to transfer "learning to adapt" information across datasets.

## 5. Conclusion

In this paper, we proposed a novel Parsing Induced Learner (PIL) to assist human pose estimation by effectively exploiting parsing information. PIL learns to predict certain pose model parameters from parsing features and adapts the pose model to extracting complementary useful features. The whole model is end-to-end trainable. Comprehensive experiments on single- and multi-person pose estimation benchmarks LIP and extended PASCAL-Person-Part demonstrated advantages of the proposed PIL over other parsing utilization approaches, including traditional multi-task learning. In addition, cross-dataset evaluation by utilizing PIL trained on LIP dataset to MPII dataset showed the PIL offers appealing transferability. Even if the applied dataset does not provide any parsing information, externally pre-trained PIL still helps the model achieve new state-of-the-art.

## Acknowledgement

Jiashi Feng was partially supported by NUS startup R-263-000-C08-133, MOE Tier-I R-263-000-C21-112, NUS IDS R-263-000-C67-646 and ECRA R-263-000-C87-133.

## References


[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 5

[2] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi. Learning feed-forward one-shot learners. In *NIPS*, 2016. 3, 4

[3] H. Bilen and A. Vedaldi. Integrated perception with recurrent multi-task neural networks. In *NIPS*, 2016. 2

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2, 5

[5] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *ICCV*, 2016. 7

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentatin with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 5

[7] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 2

[8] X. Chen and A. L. Yuille. Parsing occluded people by flexible compositions. In *CVPR*, 2015. 7

[9] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *ICCV*, 2017. 2, 7, 8

[10] C.-J. Chou, J.-T. Chien, and H.-T. Chen. Self adversarial training for human pose estimation. In *CVPR Workshop*, 2017. 2, 6, 7, 8

[11] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *CVPR*, 2017. 1, 7

[12] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 2

[13] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014. 1, 2, 3

[14] H. Fang, S. Xie, Y. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2016. 2

[15] G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. In *ECCV*, 2016. 7

[16] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 2, 5, 6

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[18] P. Hu and D. Ramanan. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *CVPR*, 2016. 7

[19] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. 5, 7

[20] L. Ladicky, P. H. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *CVPR*, 2013. 1, 2, 3

[21] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015. 1, 2

[22] I. Lifshitz, E. Fetaya, and S. Ullman. Human pose estimation using deep consensus voting. In *ECCV*, 2016. 7

[23] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan. Matching-cnn meets knn: Quasi-parametric human parsing. In *CVPR*, 2015. 1, 2

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[25] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1, 2, 3, 5, 6, 7

[26] X. Nie, J. Feng, J. Xing, and S. Yan. Generative partition networks for multi-person pose estimation. *arXiv preprint arXiv:1705.07422*, 2017. 2, 5, 7

[27] A. Paszke, S. Gross, and S. Chintala. Pytorch, 2017. 5

[28] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV*, 2013. 7

[29] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 5, 7

[30] U. Rafi, B. Leibe, J. Gall, and I. Kostrikov. An efficient convolutional network for human pose estimation. In *BMVC*, 2016. 7

[31] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 5

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 3, 5

[33] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012. 5

[34] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *CVPR*, 2015. 7

[35] 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 *NIPS*, 2014. 7

[36] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 1, 2, 5, 7

[37] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human part segmentation with auto zoom net. In *ECCV*, 2015. 1, 2

[38] F. Xia, P. Wang, X. Chen, and A. Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, 2017. 1, 2, 3, 5, 7

[39] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 1, 2, 3

[40] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *ICCV*, 2017. 1, 2, 7

[41] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2878–2890, 2013. 5

[42] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 5