Human Pose Estimation from a Single Image

Yikai Wang
Student id: 2020233280

Chuanhao Hao
Student id: 2020233176

Abstract

Convolutional networks are powerful visual models that yields hierarchies of features. In this paper, we improve the baselines models for human pose estimation based on ResNet[3], we use fully convolutional layers to replace the deconvolutional layers in origin models.The experiment results demonstrate that our method still have a good performance.

1. Introduction

Human pose estimation, which has been extensively studied in computer vision literature, involves estimating the configuration of human body parts from input data captured by sensors, in particular images and videos. Human pose estimation provides geometric and motion information of the human body which has been applied to a wide range of applications such as human-computer interaction, motion analysis, augmented reality(AR), virtual reality(VR), healthcare,etc. With the rapid development of deep learning solutions in recent years, such solutions have been shown to outperform classical computer vision methods in various tasks including image classification, semantic segmentation and object detection. Significant progress and remarkable performance have already been made by employ deep learning techniques in human pose estimation tasks. However, challenges such as occlusion, insufficient training data, and depth ambiguity still pose difficulties to be overcome. 2D human pose estimation from images and videos with 2D pose annotations is easily achievable and high performance has been reached for the human pose estimation of a single person using deep learning techniques. 2D single-person pose estimation is used to localize human body joint positions when the input is a single-person image. A good human pose estimation system must be robust to occlusion and severe deformation, successful on rare and novel poses, and invariant to challenge in appearance due to factors like clothing and lighting. Early work tackles such difficulties using robust features and sophisticated structured prediction: the former is used to produce local interpretations, whereas the latter is used to infer a globally consistent pose.

However, this conventional pipeline has been greatly reshaped by convolutional neural networks[6], a main driver behind an explosive rise in performance across many computer vision tasks. Recently pose estimation[13] systems have universally adopted convolutional neural networks as their main building blocks, largely replacing hand-crafted features and graphical models; this strategy has yielded drastic improvements on standard benchmarks. In our project, we utilize the ResNet[3] to improve the simple baselines for pose estimation[15], we simplify the base lines method and still achieve a good result.

2. Related Work

Traditionally 2D human pose estimation method adopt different hand-crafted feature extraction techniques for body parts, and these early works describe human body as a stick figure to obtain global pose structures. Recently, deep learning-based approaches have achieved a major breakthrough in human pose estimation by improving the performance significantly.

Using AlexNet[5] as the backbone, Toshev et al[14] proposed a cascaded deep neural network regressor named DeepPose, with the introduction of "DeepPose" , research on human pose estimation began shifting from classic approaches to deep learning networks. Toshev et al use their network to directly regress the x,y coordinate of joints. The work by Tompson et al[13] instead generates heatmaps by running an image through multiple resolution banks in parallel to simultaneously capture features at a variety of scales. A critical feature of the method proposed by Tompson et al[13] is the joint use of ConvNet and a graphical model. Their graphical model learns typical spatial relationships between joints. Others have recently tackled this in similar ways[10] with variation on how to approach unary score generation and pairwise comparison of adjacent joints. Based on GoogLeNet[12], Carreira et al.[1] proposed an Iterative Error Feedback(IEF) network, which is a self-correcting model to progressively change an initial solution by injecting the prediction error back to the input space. Sun et al.[11] introduced a structure-aware regres-

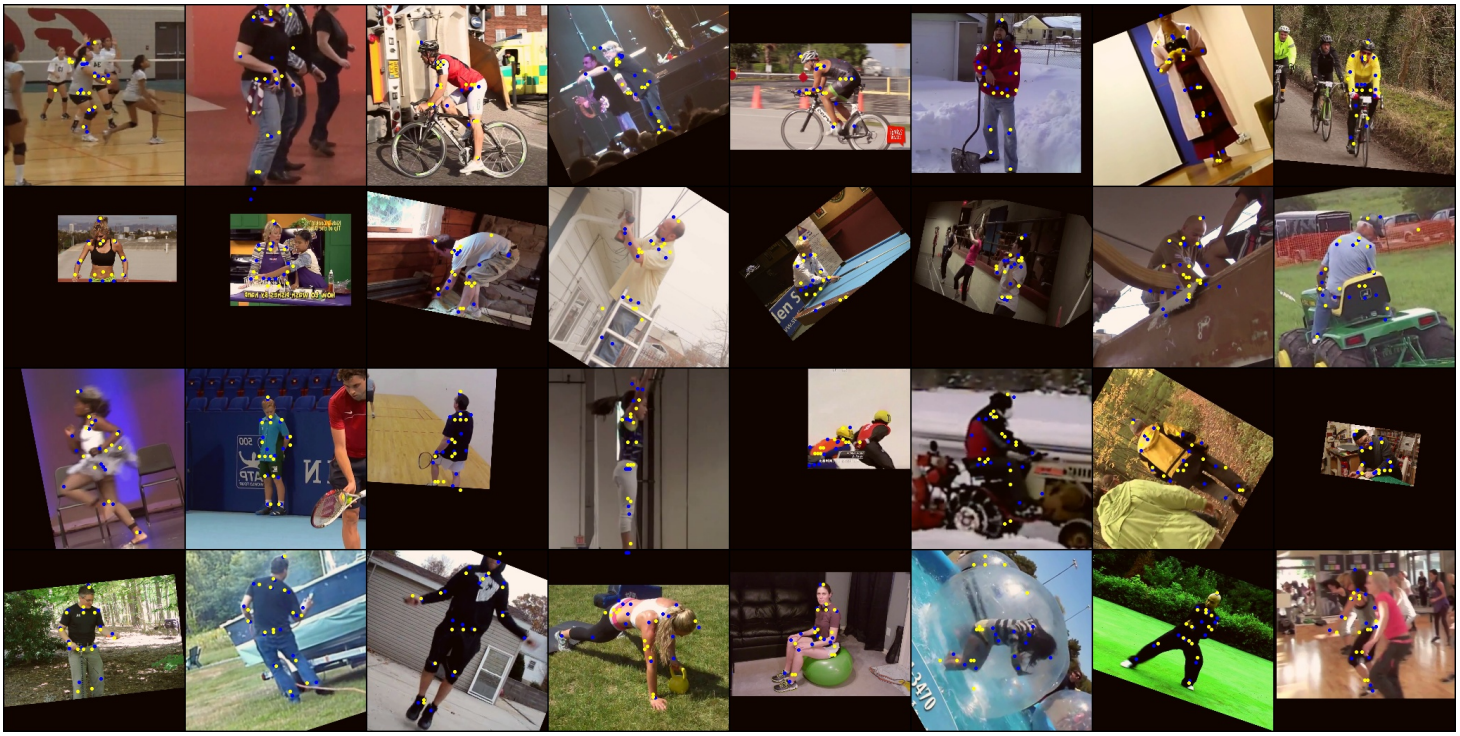

Figure 1. prediction result of our method

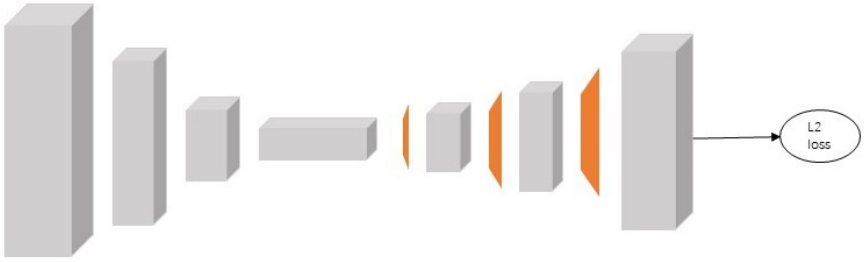

Figure 2. baseline model

sion method called ”compositional pose regression” based on ResNet-50[3]. This method adopts a re-parameterized and bone-based representation that contains human body information and pose structure, instead of the traditional joint-based representation. Luvizon et al.[8] proposed an end-to-end regression approach for human pose estimation using soft-argmax function to convert feature maps into joint coordinates in a fully differentiable framework.

A good feature that encodes rich pose information is critical for regression-based methods. One popular strategy to learn better feature representation is muti-task learning. By sharing representation between related tasks(eg., pose estimation and pose-based action recognition), the model can generalize better on the original task(pose estimation). Following this direction, Li et al[7] proposed a heterogeneous multitask framework that consists of two tasks: predicting joints coordinates from full images by building a regressor and detecting body parts from image patches using a sliding window. Fan et al.[2] proposed a Dual-Source(i.e. images patches and full images) Deep Convolutional Neural Network (DS-CNN) for two tasks: joint detection which determines whether a patch contains a body joint, and joint localization which finds the exact location of the joint in the patch. Each task corresponds to a loss function, and the combination of two tasks leads to improved results. Luvizon et al[8] learned a muti-task network to jointly handle

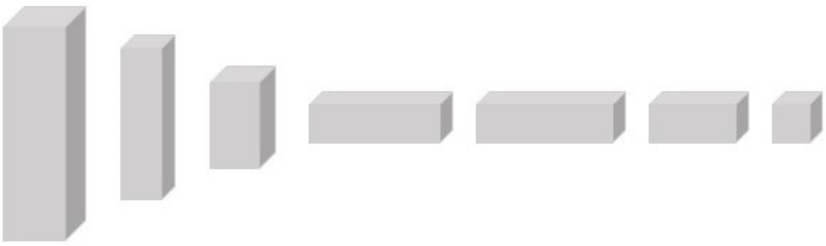

Figure 3. our method

Table 1. Results on MPII Human Pose(PCKh@0.5)

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Tompson et al.[9]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Carreira et al[9].	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Pishchulin et al.[9]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Hu et al.[9]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Wei et al[9].	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
our model	**86.39**	**65.79**	**50.01**	**33.28**	**50.56**	**37.64**	**40.15**	**52.27**

2D/3D pose estimation and action recognition from video sequences.

3. Data

Max Planck Institute for Informatics(MPII) Human Pose Dataset: This is a popular dataset for evaluation of articulated human pose estimation. The dataset includes around 25000 images containing over 40000 individuals with annotated body joints. The images were systematically collected by a two-level hierarchical method to capture everyday human activities. The entire dataset covers 410 human activities and all the images are labeled. Each image was extracted from a YouTube video and provided with preceding and following un-annotated frames. Moreover, rich annotations including body part occlusions, 3D torso and head orientations are also labeled.

Data Augmentation: For MPII dataset, it is standard to utilize the scale and center annotations provided with all images. For each sample, these values are used to crop the image around the target person. All input images are then resized to 256×256 pixels. We do data augmentation that includes rotation(+/-30 degrees), and scaling(.75-1.25). We avoid translation augmentation of the image since location of the target person is the critical cue determining who should be annotated by the network.

4. Methods

4.1. Baseline Model

This section we provided a detail description of the baseline model we use, which is the previous state-of-the-art pose estimation, this model's best results achieves the state-of-the-art at mAP of 73.7 on COCO testdev split, which has an improvement of 1.6% and 0.7% over the winner of COCO 2017 key point Challenge's single model and their ensembled model. This pose estimation is based on a few deconvolutional layers added on a backbone network, ResNet[3]. It is probably the simplest way to estimate heat maps from deep and low resolution feature maps.

4.1.1 Model Architecture

This baseline model adds a few deconvolutional layers over the last convolution stage in the ResNet. By default, three deconvolutional layers with batch normalization and ReLU activation are used. Each layer has 256 filters with 4×4 kernel. The stride is two, a 1×1 convolutional layer is added at last to generate predicted heatmaps$H_1, \ldots, H_k$ for all k keypoints. This model combines the upsampling and convolutional parameters into deconvolutional layers without using skip layer connections. Figure.2 shows the architecture of the baseline model.

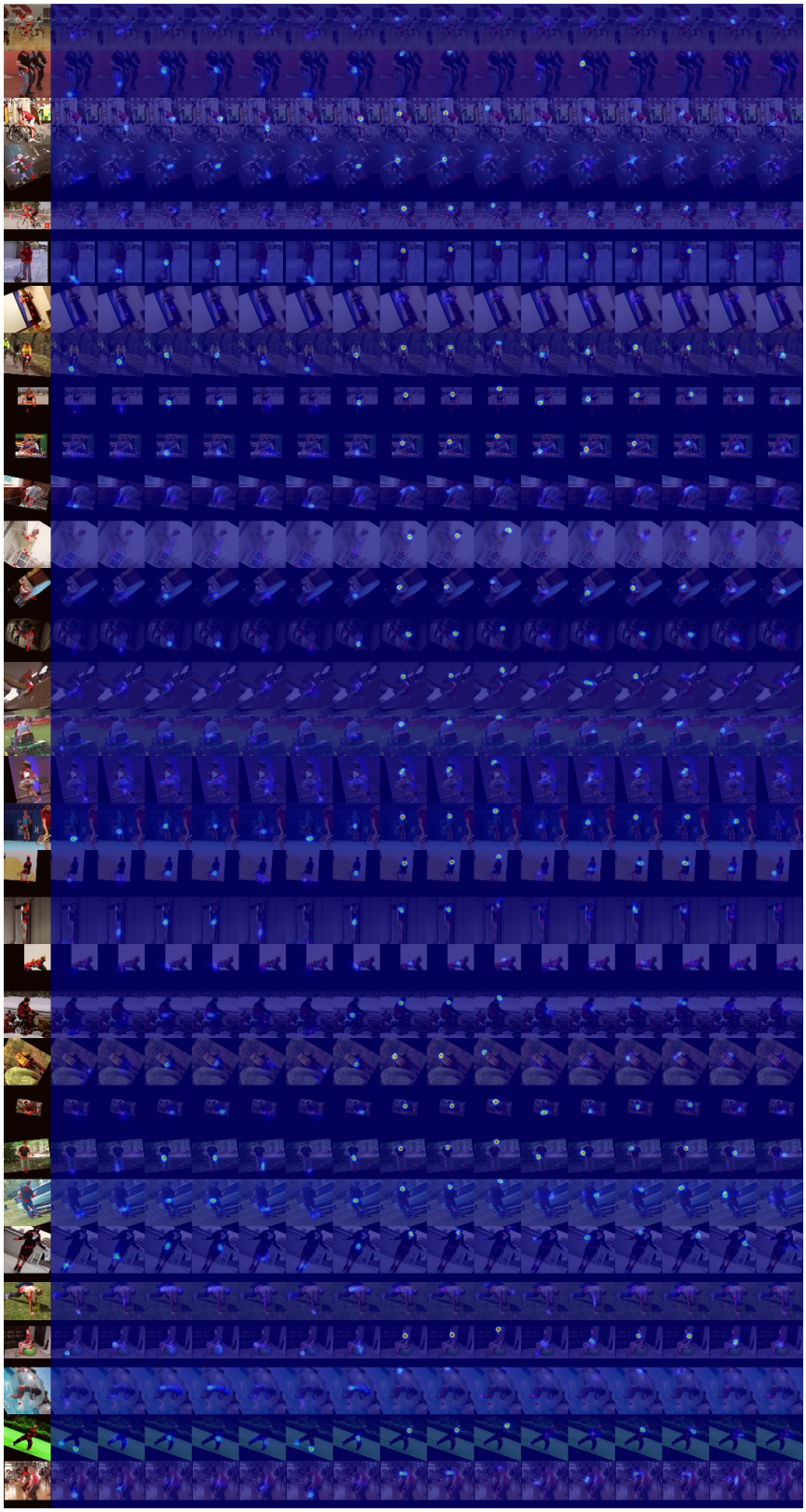

Figure 4. Example output produced by our network. On the left we see the final pose estimate provided by the max activations across each heatmap. On the right we show sample heatmaps.

4.2. Our Method

ResNet is the most common backbone network for image feature extraction, It is also used for pose estimation. Our method also simply adds a few fully convolutional networks to replace the deconvolutional layers over the last convolution stage in the ResNet, we adopt this structure because it is arguably the simplest to generate heatmaps from deep and low resolution features.

Our ResNet[3] backbone network is initialized by pretraining on the ImageNet classification. The whole network is set up as follows: Convolutional and max pooling layers are used to process features down to a very low resolution, after reaching the output resolution of the network, three consecutive rounds of 1×1 convolutions are applied to produce the final network predictions. The output of the network is a set of a joint's presence at each and every pixel. Figure.3 shows the architecture of our model, from the image, the left half model is the same as baseline model, the right half of model is different from the baseline model, our model is simpler.

4.2.1 Fully Convolutional Networks

Each layer of data in a convet is a three dimension array of size $h \times w \times d$, where h and w are spatial dimension, and d is the feature or channel dimension. The first layer is the image, with pixel size $h \times w$, and d color channels. Location in higher layers correspond to the locations in the image they are path-connected to, which are called they receptive fields.

Convets are built on translation invariance. Their basic components(convolution, pooling, and activation functions) operate on local input regions, and depend only on relative spatial coordinates. Writing $x_{i,j}$ for the data vector at location (i, j) in a particular layer, and $y_{i,j}$ for the following layer, these functions compute outputs $y_{i,j}$ by

$$y_{i,j} = f_{ks}(x_{si+\delta i,sj+\delta j 0\leq\delta i,\delta j\leq k)}$$

where k is called the kernel size, s is the stride or the subsampling factor, and f_{ks} determine the layer type: a matrix multiplication or convolution or average pooling, a spatial max for max pooling, or an elementwise nonlinearity for an activation function, and so on for other types of layers.

This function form is maintained under composition, with kernel size and stride obeying the transformation rule

$$f_{ks} \circ g_{k's'} = (f \circ g)_{k'+(k-1)s',ss'}$$

While a general deep net computes a general nonlinear function, a net with only layers of this form computes a nonlinear filter, which we call a deep filter or fully convolution network. An FCN naturally operates on an input of any size, and produces an output of corresponding(possibly resampled) spatial dimensions.

A real-valued loss function composed with FCN defines a task. If the loss function is a sum over the spatial dimensions of the final layer, $l(x, \theta) = \sum l'(x_{ij}; \theta)$, its gradient will be a sum over the gradients of each of its spatial components. Thus stochastic gradient descent will be the same as stochastic gradient descent on l', taking all of the final layer receptive fields as a minnibatch.

When these receptive fields overlap significantly, both feedforward computation and backpropagation are much more efficient when computed layer-by-layer over an entire image instead of independently patch-by-patch.

5. Experiments

5.1. Training Details

We evaluate our network on one datasets, MPII Human Pose. It consists of around 25k images with annotations for multiple people providing 40k annotated samples(28k training, 11k testing). The test annotations are not provided so in all of our experiments we train on a subset of training images while evaluating on a heldout validating set of around 3000 samples. MPII consists of images taken from a wide range of human activities with a challenge array of widely articulated full-body poses. There are often multiple people visible in a given input image, but without a graphical model or other post processing step the image must convey all necessary information for the network to determine which person deserves the annotation. We deal with this by training the network to exclusively annotate the person in the direct center.

The network is trained using Torch and for optimization we use rmsprop with a learning rate of 2.5e-4. We drop the learning rate once by a factor of 5 after validation accuracy plateaus. Batch normalization[4] is also used to improve training. We also add a batch normalization to every down sample layer. For generating the final test predictions we run both the original input and a flipped version of the image through the network and average the heatmaps together. The final prediction of the network is the max activating location of the heatmap for a given joint.

The same techniques as Tompson et al.[13] is used for supervision. A Mean-Square-Error(MSE) loss is applied comparing the predicted heatmap to a ground-truth heatmap consisting of a 2D gaussion(with standard deviation of 1 px) centered on the joint location. To improve performance at high precision thresholds the prediction is offset by a quarter of a pixel in the direction of its next highest neighbor before transforming back to the original coordinate space of the image. In MPII Human Pose, some joints do not have a corresponding ground truth annotation. In these cases the joint is either truncated or severely occluded, so far supervision a ground truth heatmap of all zeros is provided.

5.2. Results

Fig.3 shows the prediction result of our method, according to the image, our method can predict the joints approximately. The blue points are our predicted points, the yellow points are the ground truth points.

5.2.1 Evaluation

Evaluation is done using the standard Percentage of Correct Keypoints(PCK) metric which reports the percentage of detections that fall within a normalized distance of the ground truth. Table 1 shows the comparsion with other methods on MPII Human Pose, note that the results in Table 1 are cited from [9] and not implemented by us. Therefore, the performance difference could come from implementation difference. Due to the limit of computational resource, we only train our model less than 100 epochs, Nevertheless, we believe it is safe to conclude that our baseline has comparable results but is simple.

5.3. Further Analysis

5.3.1 Multiple People

The issue of coherence becomes especially important when there are multiple people in an image. The network has to decide who annotate, but there are limited options for communicating who exactly deserves the annotation. For the purpose of this work, the only signal provided is the centering and scaling of the target person trusting that the input will be clear enough to parse. Since we are training a system to generate pose predictions for a single person, the idea output in an ambiguous situation would demonstrate a commitment to the joints of just one figure. Even if the predictions are lower quality, this would show a deeper understanding of the task at hand. Estimating a location for the wrist with a disregard for whom the wrist may belong is not desired behavior from a pose estimation system.

A more comprehensive management of annotations for multiple people is out of the scope of this work. Many of the system's failure cases are a result of confusing the joints of multiple people, but it is promising that in many examples with severe overlap of figures the network will appropriately pick out a single figure to annotate.

5.3.2 Occlusion

Occlusion performance can be difficult to assess as it often falls into two distinct categories. The first consists of cases where a joint is not visible but its position is apparent given the context of the image. MPII generally provides ground truth locations for these joints, and an additional annotation indicates their lack of visibility. The second situation, on the other hand, occurs when there is absolutely no information about where a particular joint might be. For example, images where only the upper half of the person's body is visible. In MPII these joints will not have a ground truth annotation associated with them.

Our model makes no use of the additional visibility annotations, but we still makes strong estimates in most cases. In many examples, the network prediction and ground-truth annotation may not agree while both residing in valid locations, and the ambiguity of the image means there is no way to determine which one is truly correct.

We also consider the more extreme case where a joint may be severely occluded or truncated and therefore have no annotation at all. The PCK metric used when evaluating pose estimation systems does not reflect how well these situations are recognized by the network. If there is no ground truth annotation provided for a joint it is impossible to assess the quality of the prediction made by the system, so it is not counted towards the final reported PCK value. Because of this, there is no harm in generating predictions for all joints even though the predictions for completely occluded or truncated joints will make no sense. For use in a real system, a degree of meta knowledge is essential, and the understanding that no good prediction can be made on a particular joint is very important. We observe that our network gives consistent and accurate predictions of whether or not a ground truth annotation is available for a joint.

We consider the ankle and knee for this analysis since these are occluded most often. Lower limbs are frequently cropped from images, and if we were to always visualize all joint predictions of our network, example pose figures would look unacceptable given the nonsensical lower body predictions made in these situations. For a simple way to filter out these cases we examine how well one can determine the presence of an annotation for a joint given the corresponding heatmap activation. We consider thresholding on either the maximum value of the heatmap or its mean.

6. Conclusions

In conclusion, our method simplify the baseline model without drop too much accuracy. Due to the limited of computation resource, we do not test more layers and deeper networks, in the future, we may carry out the following ideas:

1. Apply encoder-decoder layers into our model instead of simple fully connected convolutional layers.

2. Adopt a GAN to train the model which may be better at helping network learn features that are more robust.

References

[1] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iter-

ative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016. 1

[2] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1347–1355, 2015. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2, 3, 5

[4] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. arxiv e-prints. 2015. 5

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks mark. *Commun. ACM*, 60(6):84–90, 2017. 1

[6] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[7] Sijin Li, Zhi-Qiang Liu, and Antoni B Chan. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 482–489, 2014. 2

[8] Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, 2019. 2

[9] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 3, 6

[10] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016. 1

[11] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017. 1

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[13] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27:1799–1807, 2014. 1, 5

[14] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 1

[15] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1