Tiny People Pose

Lukáš Neumann and Andrea Vedaldi

Department of Engineering Science, University of Oxford
{lukas,vedaldi}@robots.ox.ac.uk

Abstract. While recent progress in pose recognition has been impressive, there remains ample margin for improvement, particularly in challenging scenarios such as low resolution images. In this paper, we consider the problem of recognizing pose from tiny images of people, down to 24px high. This is relevant when interpreting people at a distance, which is important in applications such as autonomous driving and surveillance in crowds. Addressing this challenge, which has received little attention so far, can inspire modifications of traditional deep learning approaches that are likely to be applicable well beyond the case of pose recognition. Given the intrinsic ambiguity of recovering a person's pose from a small image, we propose to predict a posterior probability over pose configurations. In order to do so we: 1) define a new neural network architecture that explicitly expresses uncertainty; 2) train the network by explicitly minimizing a novel loss function based on the data log-likelihood; and 3) estimate posterior probability maps for all joints as a semi-dense subpixel Gaussian mixture model. We asses our method on downsampled versions of popular pose recognition benchmarks as well as on an additional newly-introduced testing dataset. Compared to state-of-the-art techniques, we show far superior performance at low resolution for both deterministic and probabilistic pose prediction.

1 Introduction

Interpreting images of people is an important problem in many applications of image understanding, and, as such, has received significant attention since the early days of computer vision research [9,8,30,7]. Deep learning has dramatically improved the performance of generic object detection and segmentation by methods such as Faster/Mask R-CNN [25,11], which, when applied to interpreting people in images, achieve good performance. However, models specialized for human pose recognition still perform better than such generic approaches [19,3,5], justifying research dedicated to this problem.

In this paper, we consider the problem of pose recognition, i.e. identifying in an image the location of landmark points of the human body, such as shoulders, wrists and hips. However, we do so in a setting that has not received much attention before, namely very small images of people (Fig. 1). Most approaches to pose recognition assume in fact that input images have a relatively large resolution, such that the apparent height of a person is in the order of a few hundred pixels. We consider instead the case in which a person's height is an order of magnitude smaller, down to 24 pixels high. These *tiny people* are very

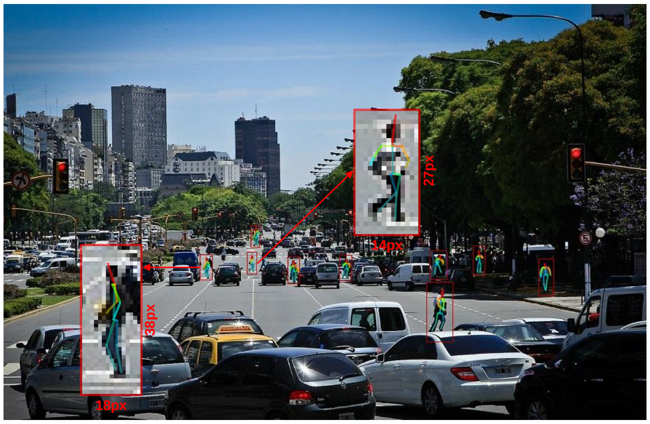


Fig. 1. Human pose estimation at a distance. Given a small low-resolution patch of an image (detected by [13]), the proposed method estimates the joint probability distribution of all body parts, despite individual pixels' noise and ambiguities. Image taken from the *Tiny People* dataset

important in applications such as surveillance and autonomous driving, where understanding the action or intention of people at a distance can be critical.

Besides the interest in applications, the tiny people problem offers unique technical challenges which inspire significant modifications of traditional deep learning approaches to structured prediction, most of which extend well beyond the case of pose recognition.

First, since there are intrinsic limitations to how much information can be extracted from small images, one should not expect to always be able to infer pose with the same degree of certainty as in the high-resolution case. Instead, we start from the outset with the goal of estimating a probability distribution over possible poses, making this distribution as tight as possible given an observed image. Doing so requires to develop a network architecture that can explicitly *express uncertainty*, both in the training as well as the testing stage.

Our second contribution is revisiting the standard "sliding window" approach for human pose estimation, in which keypoint locations are found as maxima in a dense heat map [19,5], which can be interpreted as performing a sliding window search of body parts over all image locations. However, for small images of people the resulting heat maps may have fairly low resolution. Rather than artificially increasing the resolution of the image, we consider an alternative probabilistic approach where a low resolution feature map is used to generate a dense field of Gaussian blobs, resulting in a rich *continuous* mixture model which naturally allows sub-pixel accuracy in joint prediction without significantly increasing the computational burden.

The output of the method is a distribution over possible body joint configurations, estimated from a small image of a person. We assess the method on two standard benchmarks after reducing their resolution to approximate people seen at a distance. However, this approach cannot match *exactly* the statistics of people actually imaged at a distance. For this reason, we introduce a new specialized benchmark, *Tiny People*, containing small people instances that we manually collected and annotated. While this dataset is smaller than the other standard benchmarks that we also use for assessment, we show that it is statistically sufficiently large to reliably rank algorithms and thus to provide additional verification of the validity of our approach.

On both standard benchmarks and Tiny People, we compare our method against state-of-the-art pose recognition methods. Even after carefully tuning them to the low resolution settings in different ways, we show consistently better performance in low resolution pose recognition.

For measuring performance, we consider both traditional error metrics such as average landmark regression error, as well as the *model surprise*, namely the log-likelihood of the ground-truth labels under the posterior distribution predicted by the model. The latter allows us to properly asses the quality of the probabilistic predictions. In fact, the ability of our model to express a meaningful probability prediction sets it apart from models that share some superficial technical similarity, such as [21], which lack this capability.

2 Related Work

Human Pose Estimation. The problem of estimating position of human body parts has been extensively studied before. Early methods such as the Pictorial Structures (PS) of Fischler and Elschlager [9] or the Deformable Part Models of Felzenswalb *et al.* [8] explicitly decomposed people into parts, but with the current generation of models based on deep learning the concept of parts remains implicit. Among these, Toshev and Szegedy [30] use a CNN in an iterative manner to regress human keypoint position (x_i, y_i) directly, using a L^2 loss in their training. Thompson *et al.* [29,28] adapt the FCN [18] to predict a heat map for each keypoint independently, where the keypoint position is found as position of a maximal value in the corresponding heat map - this approach has become the standard representation for human pose estimation (see Section 3.1).

With the aim of improving the network architecture by having multiple down-sample/up-sample stages, Newell *et al.* [19] introduced their Stacked Hourglass model, which is able to capture relationships at multiple scales. More recently, Chu *et al.* [5] extended Hourglass with context attention, improving its accuracy by 1 percent point on the standard MPII Human Pose dataset [1] and thus achieving the state-of-the-art result in a single person recognition.

Methods for pose estimation of **multiple people** can be divided into two subgroups: top-down and bottom-up. Top-down methods first detect bounding boxes of individual people and in a second stage a single human pose is inferred from the cropped region. Pishchulin *et al.* [24] use a model based on Pictorial Structures. He *et al.* [11] introduced Mask R-CNN for object segmentation, which is also applied to human pose estimation by predicting an individual heat map for each keypoint. Papandreou *et al.* [21] employ a Faster R-CNN [25] with ResNet-101 backbone [12] as a person detector, and a separate ResNet-101 network to process the detected bounding box to infer a 3-dimensional “heat-map” for each keypoint. The heat map predicts whether given pixel is close to the keypoint as a binary classification, and a 2-dimensional vector encoding the distance of the current pixel to the keypoint. The final keypoint position is then given by taking an average of the 2-dimensional vector for all pixels classified positively by the binary classifier.

Bottom-up methods detect individual keypoints first and then associate these parts with human instances. Insafutdinov *et al.* [14] formulate the problem of associating detected parts to human instances as a linear program over a fully-connected graph, which is an NP-hard problem. More recently, Cao *et al.* [3] introduced Part Affinity Fields, where part confidence maps together with 2D vector fields that encode position and orientation of human limbs allow greedy inference of the associations, thus making the method run in real time.

Small Objects Recognition. Several authors have focused on the task of small object recognition in a classification setting. Efros *et al.* [7] use an optical-flow motion descriptor to recognize actions of people in sports videos using a k-nearest neighbor search in a annotated people action database. Park and Ramanan [22] used low-resolution people images as feature vectors for nearest neighbor search in a database of synthetically morphed trained images to infer human pose in video sequences.

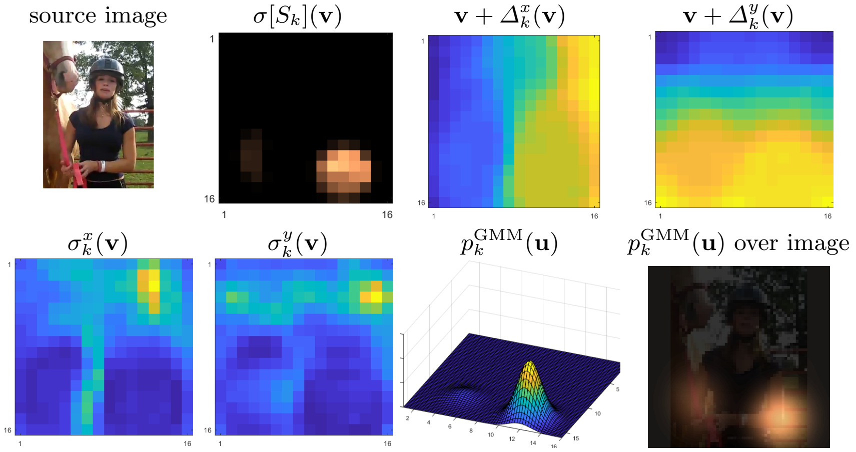
Inspirational to our work, Hu and Ramanan [13] focused on detecting tiny human faces and proposed a cascade of face detectors to detect human faces of a wide range of scales, starting with faces of 20px in height. Their task however is a region binary classification problem (i.e. telling if a bounding box contains or not a face), whereas in our work we infer 16 keypoint positions within the region of a similar size.

Modeling Uncertainty. Probabilistic models have recently started to be applied in the context of deep learning as well. Gal and Ghahramani [10] studied a relationship between Dropout [27] and Gaussian processes. Kendall and Cipolla [15] estimate uncertainty in 6-DOF camera pose estimation by Monte Carlo sampling. Novotny *et al.* [20] explicitly model uncertainty as one of the network outputs in the task of 3D geometry prediction from videos in an unsupervised setup. Most recently, Kendall and Gal [16] model both aleatoric and epistemic uncertainty within one model for dense semantic segmentation, and Rupprecht *et al.* [26] introduced a meta-loss that allows a model to learn multiple hypotheses, but do not specify how to select the best one.

3 Method

We formulate *pose estimation* as the problem of learning a function Φ mapping an image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ to a set of K-landmark locations $\mathbf{u}_1, \ldots, \mathbf{u}_K \in \mathbb{R}^2$, corresponding to different body landmarks, such as the left hip, the right ankle, and so on.

We begin by discussing the typical formulation of pose estimation (Section 3.1) and then introduce our probabilistic formulation (Section 3.2), which is able to model uncertainties arising from estimating human pose from a low resolution image.



source image

$\sigma[S_k](\mathbf{v})$

$\mathbf{v} + \Delta_k^x(\mathbf{v})$

$\mathbf{v} + \Delta_k^y(\mathbf{v})$

$\sigma_k^x(\mathbf{v})$

$\sigma_k^y(\mathbf{v})$

$p_k^{\mathrm{GMM}}(\mathbf{u})$

$p_k^{\mathrm{GMM}}(\mathbf{u})$ over image

Fig. 2. The method emits a continuous Gaussian distribution $p_k^{\mathrm{GMM}}(\mathbf{u})$ for each keypoint k (*left elbow* shown above) by estimating Gaussian Mixture Model parameters using a coarse 16×16 feature map $\mathbf{v} \in \Omega_d$ generated over the whole image. Note that the resulting distribution has in principle infinite resolution, given only by the sampling step, and that it can also express multiple modes (second but smaller mode in the location of the right elbow in the above example)

3.1 Standard formulation

The standard approach to implement a pose detector with a neural network is to express the landmark locations as maxima of a corresponding number of heat maps. To this end, let $\Phi : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{\frac{H}{\delta} \times \frac{W}{\delta} \times K}$ where K is the number of body landmarks and $\delta \geq 1$ a downsampling factor (used for efficiency). Then a landmark's location is obtained as $\mathbf{u}_k = \mathrm{argmarx}_{\mathbf{u} \in \Omega_\delta} S_k(\mathbf{u}; \Phi\mathbf{x})$, where $\Omega_\delta = \{1, 2, \ldots, H/\delta\} \times \{1, 2, \ldots, W/\delta\}$ is the downscaled version of the image domain and we extracted the k-th heat map $S_k(\mathbf{u}; \Phi\mathbf{x}) = [\Phi(\mathbf{x})]_{(\mathbf{u},k)}$ from the deep network's output. We implicitly assume that all landmarks coordinate are expressed relatively to that coordinate frame.

During training, the heat maps must be fitted to ground-truth data so that they strongly respond to the correct location of the landmarks. Very often this is done by direct L^2 regression of a manually-defined heat map shape. That is, if $\mathbf{u}_k^*$ is the ground truth location of a landmark, one minimizes the loss

$$\mathcal{L}_1(\Phi) = \sum_k \sum_{\mathbf{u} \in \Omega_\delta} w(\mathbf{u} - \mathbf{u}_k^*) \left(S_k(\mathbf{u}; \Phi\mathbf{x}) - g(\mathbf{u} - \mathbf{u}_k^*) \right)^2 \tag{1}$$

where g is a Gaussian-like kernel and w a weighting function which encourages g to be fitted well around its maximum (as otherwise the minimization would be dominated by the part of the domain far away from the peak, which occupy the majority of the image area).

Note that, while this method can successfully learn heat maps with maxima at the desired location, the choice of loss (1) and the Gaussian-like kernel g parameters is purely heuristic. Furthermore, the heat map does not convey any information beyond the location of its maximum. In the next section, we remedy this situation by introducing a probabilistic model instead, where heat maps become probability densities.

3.2 Probabilistic formulation

It is possible to explicitly turn a heath map in *posterior probability over possible locations* of a certain landmark, for example by means of the softmax operator $\sigma[\cdot]$:

$$p_k(\mathbf{u}|\mathit{\Phi}\mathbf{x}) = \sigma[S_k](\mathbf{u}) = \frac{\exp S_k(\mathbf{u};\mathit{\Phi}\mathbf{x})}{\sum_{\mathbf{v}\in\mathit{\Omega}} S_k(\mathbf{v};\mathit{\Phi}\mathbf{x})}$$

The joint probability of combined landmark locations can then be written as the product:

$$p(\mathbf{u}_1,\dots,\mathbf{u}_K|\mathit{\Phi}\mathbf{x}) = \prod_k p_k(\mathbf{v}_k|\mathit{\Phi}\mathbf{x}).$$

With this definition, loss (1) is replaced by the model likelihood, i.e. the probability of the observations $(\mathbf{u}_1^*,\dots,\mathbf{u}_K^*)$ under the model:

$$\mathcal{L}_2(\mathit{\Phi}) = -\sum_k \log p_k(\mathbf{u}_k^*|\mathit{\Phi}\mathbf{x}). \tag{2}$$

This is a simple but powerful change because it allows the model to properly represent uncertainty in a landmark's location.

In order to efficiently encode high-resolution pose information, we consider a further extension of this architecture. We assume in particular that the feature map has fairly low resolution, so that it is efficient to compute. We also assume that the network estimates at each location $\mathbf{v}\in\mathit{\Omega}_\delta$ a displacement $\mathit{\Delta}_k(\mathbf{v};\mathit{\Phi}\mathbf{x})\in\mathbb{R}^2$ and a covariance matrix $\mathit{\Sigma}_k(\mathbf{v};\mathit{\Phi}\mathbf{x})\in\mathbb{S}_+^2$ In this manner, each feature map location *emits a Gaussian distribution* over possible landmark locations. This results in the mixture model:

$$\hat{p}_k(\mathbf{u}|\mathit{\Phi}\mathbf{x}) = \sum_{\mathbf{v}\in\mathit{\Omega}_d} \mathcal{N}(\mathbf{u}|\mathbf{v}+\mathit{\Delta}_k(\mathbf{v};\mathit{\Phi}\mathbf{x}),\mathit{\Sigma}_k(\mathbf{v};\mathit{\Phi}\mathbf{x}))\cdot p_k(\mathbf{v}|\mathit{\Phi}\mathbf{x}). \tag{3}$$

A key advantage of this model is that, while the summation is carried over a discrete (and coarse) domain, the resulting distribution p_k^{GMM} is continuous and thus has, in principle, infinite resolution. This can be interpreted as a probabilistic method to perform sub-pixel interpolation.

This probabilistic model is learned by maximizing the posterior log-probability of the ground-truth observations. Hence, given a training dataset $\left(\mathbf{x}_i,\mathbf{u}_1^{(i)},\dots,\mathbf{u}_K^{(i)}\right)$, one optimizes the loss

$$\mathcal{L}_3(\mathit{\Phi}) = -\frac{1}{N}\sum_{i=1}^{n}\log\prod_k \hat{p}_k^{(i)}(\mathbf{u}_k^{(i)}|\mathit{\Phi}\mathbf{x}_i), \tag{4}$$

Table 1. Comparison of models' accuracy on the MPII Human Pose validation set with decreasing person height (PCKh @ 0.5)

method	input dimension	256px	128px	96px	64px	48px	32px	24px
HG [19]	256×256	**83.5**	**81.3**	80.2	72.4	59.4	32.8	14.1
HG-DOWNUP [19]	256×256	75.7	78.3	78.6	76.2	68.7	41.0	20.0
HG-64 [19]	64×64	77.6	77.5	76.7	74.3	65.9	45.2	18.9
Part Affinity Fields [3]	256×256	65.6	65.8	65.1	61.8	57.1	50.3	31.3
our model	64×64	81.5	**81.3**	**80.5**	**79.7**	**75.0**	**61.8**	**48.0**

where $\mathbf{x}_i$ denotes the image and $\mathbf{u}_1^{(i)}, \ldots, \mathbf{u}_K^{(i)}$ denote the keypoint annotations for the i-th image. The loss (4) is differentiable w.r.t. $\Delta_k^{(i)}$, $\Sigma_k^{(i)}$ and $S_k^{(i)}$, so it can be minimized using standard gradient descent algorithms.

In our model, Φ is a fully convolutional neural network [18], which takes an input image $\mathbf{x}$ and outputs a 5-dimensional feature tensor for each keypoint k – the Gaussian mixture weights $S_k(\mathbf{v})$ and 4 Gaussian distribution parameters – means in x,y direction $\Delta_k(\mathbf{v})$ and a diagonal covariance matrix $\Sigma_k(\mathbf{v})$ for every $\mathbf{v}$ in the feature map Ω_d (see Fig. 2). We build on the Spatial ConvNet [23] architecture which is based on the popular VGG16 [4], by taking its first 8 convolutional layers and changing the input dimension to 64×64 pixels, which results in the output feature map resolution of $16 \times 16 \times 5K$.

Discussion. Equation (3) might be reminiscent of the pose estimation in the wild method [21], which also regresses a dense displacement field in order to predict landmarks. The key difference is that our formulation estimates an actual joint probability distribution and, as shown in eq. (4), optimizes over the label likelihood. While this might seem a minor difference, in reality it has a major effect. Instead of using heatmaps as a technical device to encode mere 2D positions as done in [21], our probability maps properly encode aleatoric uncertainty. By allowing the model to predict its own uncertainty we achieve more robust learning as well as a meaningful (verifiable) confidence score together that can be used in applications. In Section 4 we demonstrate both effects empirically in ablation studies.

4 Experiments

This section thoroughly evaluates our method against alternatives for low-resolution pose estimation. After discussing the learning setup (Section 4.1), the method is evaluated on two standard benchmarks — since the benchmarks do not contain small people, we scale images down in both datasets (Section 4.2). Additionally, we collected a new ad-hoc dataset of people seen at a distance, which is used only for testing the model (Section 4.3). We also show model performance in conjunction with a tiny people detector. An ablation study, which evaluates the main design choices of the proposed model, is presented in Section 4.4.

4.1 Training details

In order to train our model, we combined the MPII Human Pose dataset [1] with the MS COCO dataset [17] and resized all instances of people using bilinear

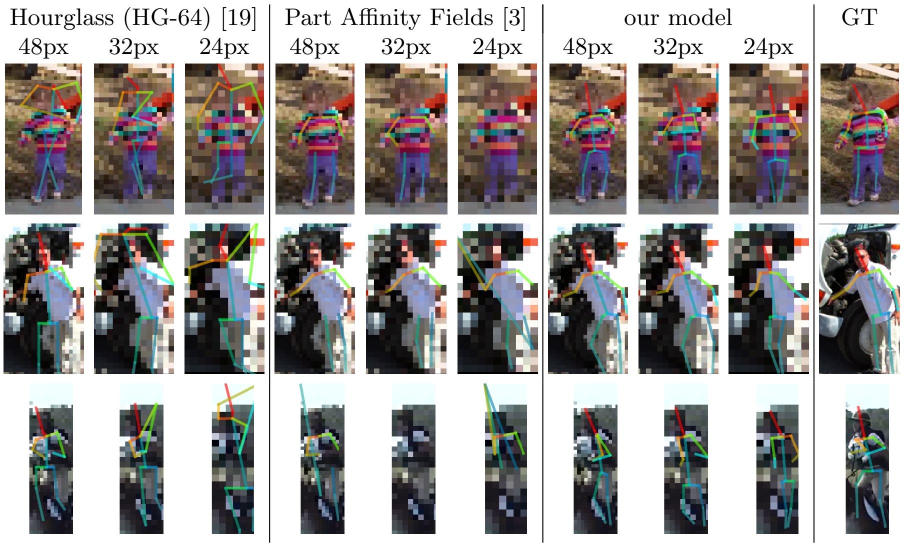
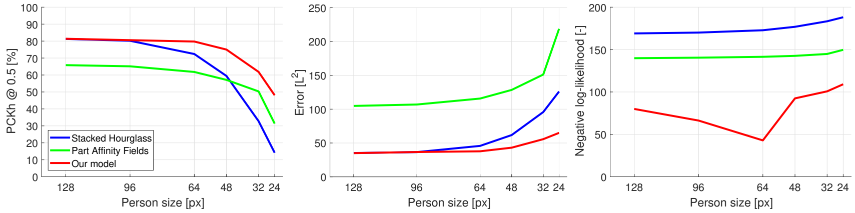


Fig. 3. Qualitative results on the **MPII Human Pose** validation set downsampled to 48, 32 and 24 pixels in height. Note that for Part Affinity Fields we only show the 12 keypoints consistent with MPII Human Pose annotation format. Note that *all methods were carefully optimized to give their best in the low-resolution setting* (see text). Best viewed zoomed in color



(a) (b) (c)

Fig. 4. Comparison of models' keypoint detection accuracy (a), regression error (b) and negative log-likelihood of the ground truth - "surprise" (c) on the **MPII Human Pose** validation set with decreasing person height

interpolation so that the resulting bounding box is 64 pixels in height (while we train the system at this resolution, we test it on much smaller images as well).

Since the testing set for the MPII Human Pose dataset is not publicly available, we split the MPII training set into training and validation subsets, following the data split published by Thompson *et al.* [29] and used elsewhere in the literature [2,31]. Note that because the MPII and COCO datasets slightly differ in the number of human keypoints annotated (16 vs 18), we trained our model to predict 16 keypoints, following the MPII Human Pose dataset format.

We trained a single model using both training subsets of MS COCO and MPII for 80 epochs using vanilla SGD, using the learning rate 10^{-5} for the first 40 epochs and then dropping the rate to a half every 10 epochs. For data augmentation, we only used a random scale augmentation in the range $(0.7, 1.3)$. In the training, we also constrain the displacement values $|\Delta(\mathbf{v}; \Phi\mathbf{x})| \leq 3$ (measured in the final $16x16$ feature layer), because this makes the training process more numerically stable - this constraint effectively ensures that the individual elements in the feature map $\mathbf{v} \in \Omega_d$ do not contribute to keypoint locations which are physically outside of their individual receptive fields.

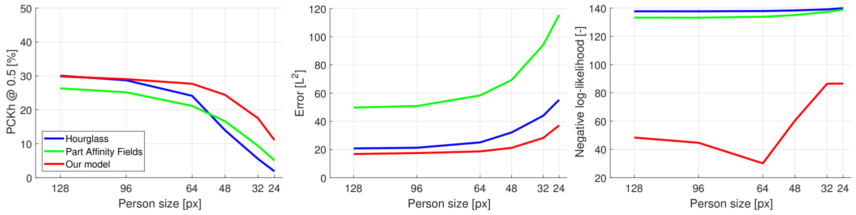
4.2 Standard Benchmarks

We assess our approach using two popular standard human pose benchmarks, MPII Human Pose [1] and MS COCO [17]. Images are downsampled to create people of predefined sizes, starting from 256 pixels and going down to 24 pixels in height, in order to evaluate the impact of resolution to pose estimation accuracy.

Baselines. We compare against two state-of-the-art methods for human pose estimation: Stacked Hourglass (HG) [19] and Part Affinity Fields [3]. We do our best to maximize the performance of these methods on our low-resolution data and test three approaches: naïve, retraining, and architecture editing. The **naïve** approach is to use off-the-shelf models trained by the respective authors and simply upscale the input using bilinear interpolation to the resolution expected by the model (256×256 pixels, which requires up to 8× upsampling). By comparison, input images to our model area always down- or up-sampled to 64×64 pixels using the same mechanism. Since upsampling may change the image statistics, the second approach is to **retrain** the Stacked Hourglass model from scratch on the MPII dataset, using the training code and default parameters provided by the authors, on images first downsampled to 64×64 pixels and then upsampled to the expected resolution on 256×256 (HG-DOWNUP). The third approach is to **edit** the architecture to work directly on the same low-resolution data as ours. For Hourglass, we do so by removing the first two max-pooling layers to natively work on 64×64 input patches and retrain the model using the same protocol (HG-64). To make the comparison fair, we also include the results of our method when only trained on the MPII dataset (MPI-TRAIN).

Results. On the MPII Human Pose dataset (see table 1 and Fig. 3), for people 128px high our method achieves competitive accuracy to the naïve Stacked Hourglass (HG) model [19], even though our effective input image size is in fact 64px, and outperforms Part Affinity Fields by a large margin. At 64px our model outperforms Stacked Hourglass by 10 and Part Affinity Fields (PAFs) by 20 percent points. Finally at 24px, our model has almost two times higher keypoint detection accuracy than PAFs and four times higher accuracy than Stacked Hourglass (see Fig. 4). Note that since PAFs output keypoints in the MS COCO format, we only evaluated the 12 keypoints which can mapped to the MPII annotations format. Also note, that retraining or editing the Hourglass architecture for the low resolution data does not bring a significant boost in performance for small people (see Fig. 4).

On the MS COCO dataset (using only the 12 keypoints for all methods), our method performs on par with Stacked Hourglass for larger people (128px), but with decreasing people size the margin grows in favor of our method, significantly outperforming both the Part Affinity Fields [3] and Stacked Hourglass for people sizes smaller than 48 pixels (see Fig. 5). Compared to MPII, accuracy on MS



(a) (b) (c)

Fig. 5. Comparison of models' keypoint detection accuracy (a), regression error (b) and negative log-likelihood of the ground truth (model surprise) (c) on the **MS COCO** validation set with decreasing person size

COCO is generally worse as the data is more challenging, particularly due to partial occlusions.

In all experiments, the standard PCKh@0.5 metric [1] was used to measure keypoint detection accuracy, which requires a detection to be within certain distance from the ground truth keypoint position, where the thresholding distance is given by 50% of the head segment length.

Additionally, we calculated the L^2 regression error as the distance between the detection and the ground truth normalized by the image height, and the negative log-likelihood of the ground truth, by taking the log-likelihood $-\log p(y_{\text{gt}}|x_{\text{gt}})$ of the probabilities for the ground truth keypoint locations produced by the model for the ground truth images. This is also known as "surprise" and is an indication of the quality of the probabilistic output of the model. Since the baseline models do not output probabilities, we assumed Gaussian distribution of their output, where the mean is the predicted landmark location and the variance is constant. The actual prediction variance value was calculated for each joint by comparing predictions to ground truth in the training set. In this case, the advantage of our probabilistic model is very significant at all resolution levels, indicating that meaningful probability maps are learned.

4.3 Tiny People

A limitation of the results in Section 4.2 is that images are resized synthetically. While this is simulates well the actual application scenario which is spotting people at a distance, we cannot guarantee that the image statistics match exactly.

For this reason, we introduce a new *Tiny People* dataset.[1] The dataset was collected by searching for 21 activity categories on Flickr (the category names were taken from the MPII Human Pose dataset [1]) and manually selecting images which contain small people. We intentionally chose manual image selection to avoid a bias towards a specific detection method, so that detection accuracy can also be also evaluated.

We collected 200 images with 585 people instances (see Fig. 6), where for each person instance we annotated 14 keypoints using the open source VGG Image

[1] The dataset can be downloaded at `http://www.robots.ox.ac.uk/~vgg/data/tinyPeople/`

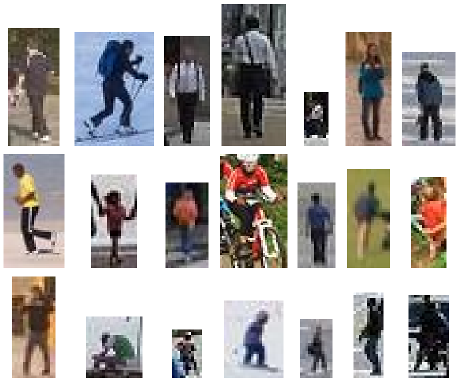


Fig. 6. Sample images from the Tiny People dataset

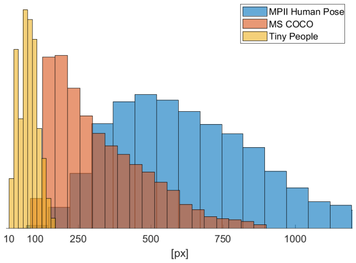


Fig. 7. Normalized histogram of people heights (as given by their bounding box annotation) in the standard datasets and in the newly introduced Tiny People dataset

Table 2. Comparison of models' accuracy on the Tiny People dataset

	PCKh @ 0.5	Avg. error [px]
Stacked Hourglass [19]	43.7	6.63
Part Affinity Fields [3]	25.5	18.99
our model	**59.9**	**4.66**

Annotator tool [6]. We followed the annotation format of the MPII Human Pose dataset (16 keypoints), but we did not annotate the *lower neck* and *pelvis* keypoints as they cannot be realistically distinguished for small people. The average person height in the dataset is 51 pixels, which is significantly lower than the existing datasets (see Fig. 7).

Pose Estimation. We again compared our model to the two state-of-the-art methods (see Fig. 8) on Tiny People. We follow the same protocol as in the previous section with images being upsampled to 256 × 256 pixels by bilinear interpolation for the existing methods. Our method achieves the accuracy of almost 60% correctly localized keypoints (see Table 2), despite being trained only on the standard datasets, which is 17 percent points higher than Stacked Hourglass [19] and more than two times better than Part Affinity Fields [3].

The dataset is of course much smaller than other benchmarks, but its size is sufficient for evaluation purposes. In particular, with 585 pose instances, the measured error standard deviation $\frac{\sigma}{\sqrt{n}}$ is 0.3px vs the error differences between methods of > 2px, which is therefore well above significance (p-value 10^{-4}). Hence a dataset of this size is sufficient to reliably rank different algorithms.

The main failure mode is confusing left and right sides, which is given by intrinsic ambiguity of human body at a small scale — when the face is not distinguishable and there is not enough context in the image, it is not clear whether the person faces towards or away from the camera and it is thus not clear which arm (leg) is left or right.

Overall, these results further support the conclusion that our model performs better on small people than the state-of-the-art methods.

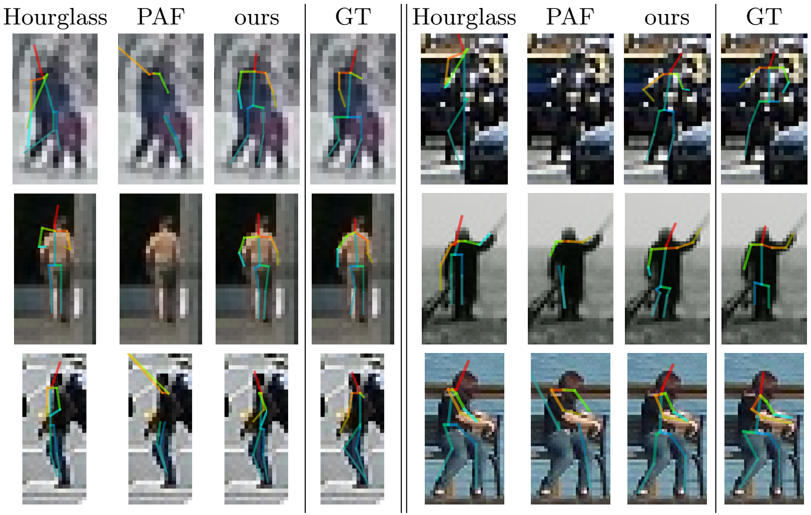


Fig. 8. Qualitative results on the **Tiny People** dataset, comparing our method against Hourglass (HG-64) [19] and PAF [3]. Note that *all methods were carefully optimized to give their best in the low-resolution setting* (see text). Best viewed zoomed in color.

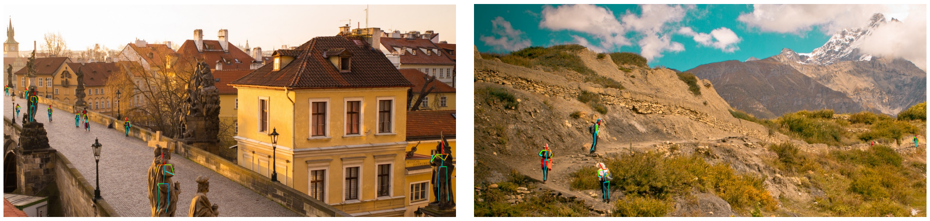


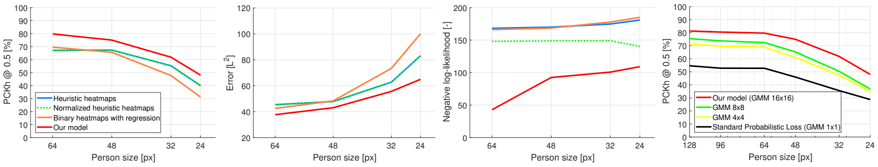
Fig. 9. Qualitative results of detection and pose estimation from the **Tiny People** dataset. Best viewed zoomed in color

Tiny People Detection. In order to show, that the proposed pose estimation method can be easily incorporated into an end-to-end pipeline, we adapted the TinyFaces [13] detector for small people detection by modifying the expected aspect ratios and scales and trained it on the MS COCO training set, where every image was down-sampled by a factor of 4 resulting in an average person height of $\approx 50px$. After detection, patches are fed as input to our method for pose estimation (see Fig. 1).

Using the standard OKS keypoint evaluation metric, the pipeline achieved the average precision $\text{AP}^{\text{OKS}=0.50}$ of 24.5% on the four times downsampled COCO Validation set (evaluating only the 12 keypoints outputted by our model) and 53.5% on the newly introduced Tiny People dataset (see Fig. 9).

4.4 Ablation study

In order to assess the impact of the newly introduced probabilistic formulation, we replace elements of our formulation with other choices common in the literature. We used the same network architecture backbone with the same input size



(a) (b) (c) (d)

Fig. 10. Comparison of different representations for tiny human pose with decreasing person size, evaluated on the **MPII Human Pose** validation set - correctly detected keypoints (a), regression error (b) and negative log-likelihood of the ground truth (c). Keypoint detection accuracy as a function of the number of the Gaussian Mixture Model elements (d)

of 64 × 64 pixels and the same training data (Section 4.1) to train models with four different representations:

- **Heuristic heatmaps**. Dense $H \times H \times K$ heat maps ($H = 32, 64, 128$) are produced by adding additional de-convolution layers to the network. Predefined Gaussian-like kernels are used to produce heat maps around each keypoint location (Section 3.1) and a standard L^2 per-pixel loss is used in the training. This setting is analogous to stacked hourglass [19] and multi-context attention [5].
- **Normalized heuristic heatmaps**. The same $32 \times 32 \times K$ heat maps as above, but normalized at test time to create a probabilistic distribution over the image space.
- **Binary heatmaps with regression**. A $16 \times 16 \times 3K$ feature map with a binary heatmap around the each keypoint location combined with a 2-D vector which encodes keypoint offset in every location of the feature space. This representation can in principle produce sub-pixel accuracy because the offset is a real-valued parameter, but it does not capture uncertainty in the training nor the testing phase. This setting is analogous to pose estimation in the wild [21].
- **Standard probabilistic loss**. Directly predicting mean and variance of every keypoint from the whole image (see Equation (2)). This is equivalent to a Gaussian mixture model with only one element.

The ablation results (see Table 3 and Fig. 10) demonstrate that the proposed probabilistic representation outperforms the standard heatmap representations by a large margin, both in terms of the keypoint detection accuracy, as well as model surprise.

Furthermore, our continuous probabilistic distribution (inferred from a feature map of 16×16 and using eq. (3)) performs better than simply increasing the heat map resolution up to 128×128 pixels by means of deconvolutional layers. In fact, doing so makes the accuracy slightly worse, probably due to over-fitting.

Also note that taking the standard heatmap representation and normalizing it to create a "probability distribution" (*Normalized heuristic heatmaps*) reduces the negative log-likelihood over assuming a Gaussian distribution (*Heuristic heatmaps* and *Binary heatmaps with regression*), but it is far inferior to the proposed approach (see Fig. 10c).

Table 3. Comparison of different representations on the MPII Human Pose validation set with decreasing person height (PCKh @ 0.5)

method	output dimension	128px	96px	64px	48px	32px	24px
Heuristic heatmaps (32x32) [19,5]	$32 \times 32 \times K$	74.5	74.4	67.1	67.3	55.2	39.9
Heuristic heatmaps (64x64)	$64 \times 64 \times K$	74.1	72.8	70.1	65.2	50.8	36.1
Heuristic heatmaps (128x128)	$128 \times 128 \times K$	74.3	72.8	71.1	65.3	49.2	35.5
Normalized heuristic heatmaps	$32 \times 32 \times K$	74.5	74.4	67.1	67.3	55.2	39.9
Binary heatmaps [21]	$16 \times 16 \times 3K$	75.5	74.5	69.6	65.6	47.7	31.2
our model (GMM 16x16)	$16 \times 16 \times 5K$	**81.3**	**80.5**	**79.7**	**75.0**	**61.8**	**48.0**
GMM 8x8	$8 \times 8 \times 5K$	75.5	73.8	72.5	65.3	50.4	36.9
GMM 4x4	$4 \times 4 \times 5K$	71.5	69.7	69.0	61.1	47.4	35.2
GMM 1x1	$4K$	54.7	52.8	52.8	46.0	35.6	28.8

Finally, we show ablation experiments for the size of the feature map $\mathbf{v} \in \Omega_d$, i.e. the number of elements in the GMM (see Fig. 10d). The model becomes worse as the number of components is reduced and on contrary adding more than $16 \times 16 = 256$ components did not improve accuracy but made the computation slower (these lines are omitted for clarity as they overlap with the others).

5 Conclusion

We have shown that low-resolution pose recognition is significantly more ambiguous than its full-resolution counterpart and that modelling uncertainty explicitly in a deep network can significantly boosts recognition performance.

We have thoroughly validated this idea. We have tested standard benchmarks resized to low resolution and a new benchmark of people seen at a distance. We have conducted numerous ablation studies that emphasize the advantage of our probabilistic approach over other common modelling choices representative of state-of-the-art methods. We have also compared our approach to such methods directly, after thoroughly optimizing them for the low resolution setting in three different ways, to demonstrate that our approach is convincingly better in the low resolution setting, showing the usefulness of the probabilistic model.

Finally, on account of the small resolution, our model runs at 250Hz on a single NVidia 1080Ti GPU, which is more than an order of magnitude faster than the existing methods, and which makes it suitable for integration to existing person detection pipelines.

Acknowledgement

We are very grateful to Continental Corporation for sponsoring this research.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Bernt, S.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)

2. Belagiannis, V., Zisserman, A.: Recurrent human pose estimation. In: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on. pp. 468–475. IEEE (2017)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
4. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014)
5. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
6. Dutta, A., Gupta, A., Zissermann, A.: VGG image annotator (VIA). http://www.robots.ox.ac.uk/ vgg/software/via/ (2016), accessed: 14.11.2017
7. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: null. p. 726. IEEE (2003)
8. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8. IEEE (2008)
9. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Transactions on computers **100**(1), 67–92 (1973)
10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059 (2016)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: The IEEE International Conference on Computer Vision (ICCV) (October 2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hu, P., Ramanan, D.: Finding tiny faces. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
14. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision. pp. 34–50. Springer (2016)
15. Kendall, A., Cipolla, R.: Modelling uncertainty in deep learning for camera relocalization. In: Robotics and Automation (ICRA), 2016 IEEE International Conference on. pp. 4762–4769. IEEE (2016)
16. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977 (2017)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
18. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
19. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. pp. 483–499. Springer (2016)
20. Novotny, D., Larlus, D., Vedaldi, A.: Learning 3d object categories by looking around them. In: The IEEE International Conference on Computer Vision (ICCV) (October 2017)

21. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild (July 2017)
22. Park, D., Ramanan, D.: Articulated pose estimation with tiny synthetic videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 58–66 (2015)
23. Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1913–1921 (2015)
24. Pishchulin, L., Jain, A., Andriluka, M., Thormählen, T., Schiele, B.: Articulated people detection and pose estimation: Reshaping the future. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 3178–3185. IEEE (2012)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
26. Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., Hager, G.D.: Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In: The IEEE International Conference on Computer Vision (ICCV) (October 2017)
27. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of machine learning research **15**(1), 1929–1958 (2014)
28. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 648–656 (2015)
29. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in neural information processing systems. pp. 1799–1807 (2014)
30. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1653–1660 (2014)
31. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4724–4732 (2016)