# Simultaneous Multi-Person Detection and Single-Person Pose Estimation With a Single Heatmap Regression Network

Christian Payer
Institute for Computer Graphics and Vision
Graz University of Technology, Austria
christian.payer@icg.tugraz.at

Thomas Neff
Institute for Computer Graphics and Vision
Graz University of Technology, Austria
thomas.neff@student.tugraz.at

Horst Bischof
Institute for Computer Graphics and Vision
Graz University of Technology, Austria
bischof@icg.tugraz.at

Martin Urschler
Ludwig Boltzmann Institute for Clinical Forensic Imaging
Graz, Austria
martin.urschler@cfi.lbg.ac.at

Darko Štern
Ludwig Boltzmann Institute for Clinical Forensic Imaging
Graz, Austria
darko.stern@cfi.lbg.ac.at

## Abstract

*We propose a two component fully-convolutional network for heatmap regression to perform multi-person pose estimation from images. The first component of the network predicts all body joints of all persons visible on an image, while the second component groups these body joints based on the position of the head of the person of interest. By applying the second component for all detected heads, the poses of all persons visible on an image are estimated. A subsequent geometric frame-by-frame tracker using distances of body joints tracks the poses of all detected persons throughout video sequences. Results on the PoseTrack challenge test set show good performance of our proposed method with a mean average precision (mAP) of 50.4 and a multiple object tracking accuracy (MOTA) of 29.9.*

## 1. Introduction

Human pose estimation from images or videos is of great interest for many applications, e.g., action recognition, sports video analytics, surveillance, and human computer interaction. Although many recent methods are capable of identifying poses of multiple persons on a single image [8, 6, 1], only few methods aim to track human poses through videos [3], which further assists the aforementioned applications.

Human pose estimation and tracking in the wild is a challenging problem due to the unconstrained settings and therefore immense variation in image appearance. To objectively compare methods, carefully annotated and publicly available datasets like PoseTrack[1] are necessary. This large-scale dataset consists of 500 video sequences with approximately 20,000 frames and 120,000 body pose annotations. As each person is annotated with a unique ID throughout the whole video sequence, this dataset allows not only evaluating human pose estimation, but also human pose tracking.

In this work, we propose a novel fully convolutional [5] network architecture using heatmap regression [10] to estimate the pose of multiple persons from a single frame. Furthermore, by employing a geometric tracker that is able to track the identified poses throughout video sequences, we show promising results on the PoseTrack dataset.

[1] https://posetrack.net

## 2. Single Frame Multi-Person Pose Estimation

Our proposed network architecture combines multiple person detection and single person pose estimation, see Fig. 1. The network architecture consists of two components: The first component performs multi-person detection of all persons visible on the image, while the second component filters these detections to predict the body joints of the current person of interest.

The outputs of the first component are the multi-person heatmaps of each body joint, as well as the multi-person heatmap of the head bounding box centers. As we are interested in the pose of each individual person, the second component needs to know, who is the current person of interest. Therefore, we detect all visible heads on the multi-person heatmap of head bounding box centers, by taking the local maxima with a minimal value $t_{\text{head}}$ and a minimal distance $t_{\text{head\_dist}}$. Then, we create a mask around the head of the current person of interest and multiply this mask with the multi-person head heatmap.

The second component takes the multi-person heatmaps of each body joint from the first component, as well as the masked head heatmap as input and filters the heatmaps to detect only the current person of interest. Thus, similar to [7], the second component can focus on filtering infeasible responses on each multi-person heatmap. An element-wise multiplication of the multi-person heatmaps from the first component with the filtered heatmaps from the second component results in the final single-person heatmaps. As each of the final single-person heatmaps contains only a single peak, we obtain the coordinate of each body joint by taking the maximum response on its single-person heatmap. To be more robust to false detections, we remove joints, whose heatmap value is below $t_{\text{joint}}$. Furthermore, we remove persons, for whom the sum of all heatmap values for all body joints is below $t_{\text{person}}$.

By processing the second network component for every head detected by the first component, the network is able to predict poses of each individual person visible on the image.

### 2.1. Network Architecture

The proposed network architecture is depicted in Fig. 1.

**Input Preprocessing and Feature Extraction:** As a first step, consecutive convolution and pooling layers act as low level feature extractors while simultaneously reducing the intermediate image resolution. The network uses RGB images with a size of $896 \times 512$ px as input, while every input image is scaled with fixed aspect ratio to fit this size. Two blocks consisting of two consecutive $3 \times 3$ convolutions with 128 outputs followed by $2 \times 2$ max pooling extract low-level features and reduce the resolution to $224 \times 128$ px.

**Multi-Person FCN:** A U-Net like sub-network takes the previously extracted features as input and creates the multi-person heatmaps. Different to the originally proposed U-Net [9], our network uses padded convolutions to keep input and output image size the same. Furthermore, we use fixed nearest-neighbor upsampling for creating the next higher level and addition instead of concatenation to combine the outputs of two levels, which reduces runtime, while keeping prediction performance high. Each intermediate convolution layer uses $3 \times 3$ kernels and creates 128 outputs. The U-Net has 6 levels with a lowest intermediate image resolution of $4 \times 7$ px resulting in a large receptive field that covers the whole image.

**Single-Person FCN:** The inputs of this U-Net like sub-network (with the same structure as the U-Net of the first component) are the extracted low-level features, the predicted multi-person heatmaps of each body joint, as well as the masked heatmap of head bounding box center of the current person of interest. The output of this network are the filtered heatmaps for the current person of interest. Multiplying these filtered heatmaps with the multi-person heatmaps from the first component results in the final single-person heatmaps.

## 3. Multi-Person Pose Tracking

The proposed geometric multi-person tracker is based on frame-by-frame matching of detected poses using minimum median distances between all body joints of detected poses. For every detected person, a new unique ID is assigned for the current frame. In order to have consistent IDs for every person over consecutive frames, we arrange tuples of persons detected in the current frame and persons detected in previous frames. For all of these tuples, we compute the median distance between every body joint from one person to the other. Starting from the minimum median distance of all tuples, if this distance is under a certain threshold $t_{\text{pose\_dist}}$, a match between a person detected in a previous frame and a person detected in the current frame is found. For this match, the newly created ID for the person detected in the current frame is set to the ID of its matching person in previous frames. Afterwards, this ID is flagged as used, in order to only allow for a one-to-one correspondence between every person detected in the current frame and all previously detected persons. This process is repeated until either all remaining combinations of tuples have distances larger than $t_{\text{pose\_dist}}$ or all previously detected persons have already been matched. All remaining IDs without matches in previous frames are assumed to be previously undetected persons. To prevent assignments of IDs for persons which have already left the visible image area, IDs are only kept for a limited amount of frames $t_{\text{max\_age}}$, afterwards they are discarded.

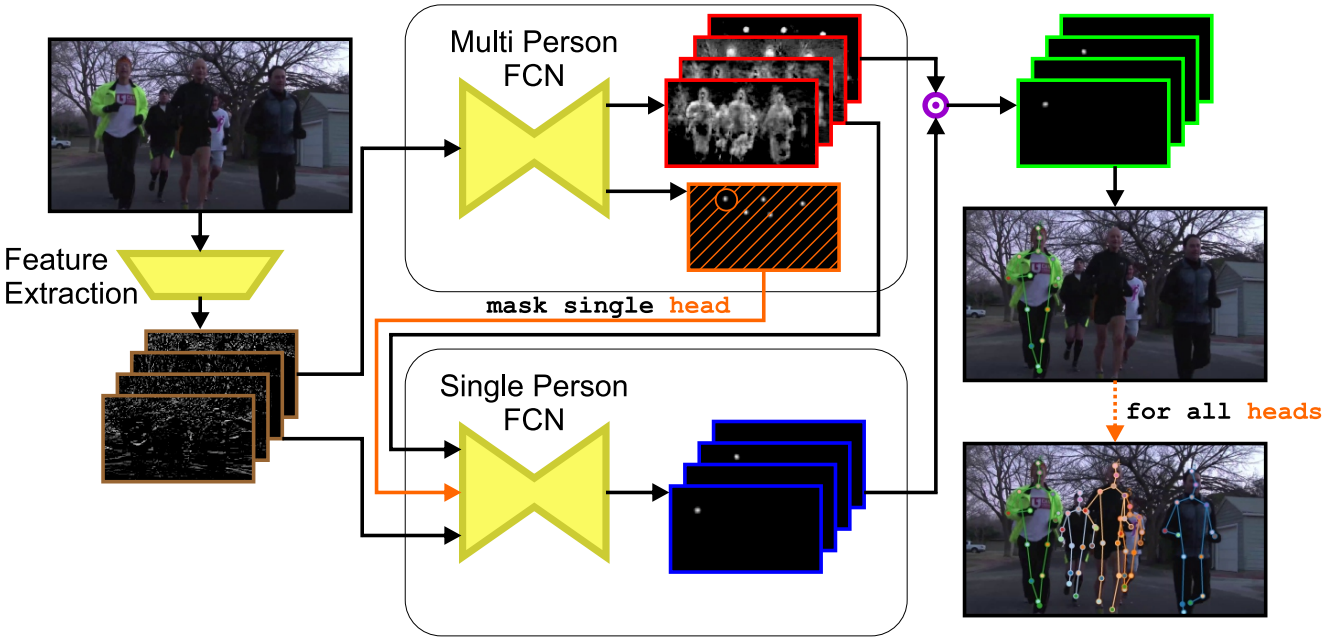


Figure 1. The proposed network architecture. Arrows denote connections; yellow boxes denote sub-networks consisting of convolution and pooling layers. The border colors of the intermediate images have the following meaning: brown: outputs from the feature extraction part; red: multi-person heatmaps of body joints; orange: multi-person heatmap of head bounding box centers; blue: filtered single-person heatmaps of body joints; green: final single-person heatmaps of body joints.

## 4. Implementation and Training Details

We evaluate our method on the PoseTrack dataset[2], which consists of 350 training, 50 validation and 100 testing video sequences. We trained and evaluated the networks in Tensorflow. The network uses RGB images as input, while we shift each intensity component by -128 and scale it by $\frac{1}{128}$ such that the range of input intensity values is [-1, 1]. In training, we perform random data augmentation. We shift the input intensity values by [-0.25, 0.25] and scale them by [0.75, 1.25]. The input images are randomly mirrored in x-axis, rotated by [-45°, 45°], scaled by [0.5, 1.5] and translated by [-100, 100]. All random transformations sample from a uniform distribution within the specified intervals.

We optimized the networks using Adam and its proposed default parameters [4] with a learning rate of 0.00005. We use a mini-batch size of 4 and optimize the networks for 1,000,000 iterations. Each intermediate convolution has a ReLU activation function, while the layers generating heatmaps have a TanH activation function. The convolution biases are initialized with 0; the weights are initialized as described in [2] and have an $L_2$ decay of 0.0005. The network loss is the sum of two $L_2$ losses, i.e., (1) the multi-person heatmap of the head bounding box centers, and (2) the single-person heatmaps (see Fig. 1).

We create the target heatmap image of the head bounding box centers as follows: Each head is represented as an image of a 2D Gaussian centered at the position of the head bounding box. The Gaussian is set to have $\sigma = 2$ px and scaled to a maximum value of 1. The target multi-person heatmap is calculated by taking the maximum over all head heatmap images of all persons on the input image. As on some images not all persons are annotated, the creators of the PoseTrack dataset supplied a pixel-wise mask of the annotated persons. To prevent overfitting due to loss calculation outside the valid region, we multiply the differences of prediction and target with this mask before calculating the loss of the head bounding box centers.

We create the target single-person heatmaps as follows: Each body joint of the person of interest is represented as an image of a 2D Gaussian centered at the position of the body joint. Each of these Gaussian images is the target single-person heatmap of the respective body joint. The parameters of the Gaussians are the same as used for the head bounding box heatmaps. In cases where we do not know the exact position of a body joint due to missing annotations, we still want the network to generate responses to prevent overfitting. Therefore, we do not care what the network predicts and do not calculate the loss for body joint heatmaps with a missing annotation.

To create the mask of head of the person of interest for the second network component, we create a Gaussian heatmap with $\sigma = 4$ px on the position of the head bounding box center. In network training, we do not perform local

[2] https://posetrack.net

| Set | Head | Sho | Elb | Wri | Hip | Knee | Ank | mAP |
|---|---|---|---|---|---|---|---|---|
| val | 73.1 | 64.2 | 56.1 | 43.0 | 48.9 | 43.7 | 39.1 | 53.9 |
| test | 66.6 | 60.6 | 51.4 | 42.4 | 43.5 | 41.7 | 38.7 | 50.4 |

Table 1. Single-frame multi-person pose estimation results showing the mean average precision (mAP).

| Set | MOTA | MOTP | Prec | Rec |
|---|---|---|---|---|
| val | 36.8 | 50.3 | 73.2 | 61.7 |
| test | 29.9 | 17.3 | 66.9 | 62.3 |

Table 2. Multi-person pose tracking results showing the multiple object tracking accuracy (MOTA) and precision (MOTP).

maxima detection of the head bounding box centers, as described in Sec. 2, but take the groundtruth position.

We obtained values of the parameters defined in Sec. 2 and 3 with grid search and set them as follows: $t_{\text{head}} = 0.2$; $t_{\text{head\_dist}} = 4$; $t_{\text{joint}} = 0.3$; $t_{\text{person}} = 3$; $t_{\text{pose\_dist}} = 200$; $t_{\text{max\_age}} = 10$.

## 5. Results

We evaluate our proposed method on the PoseTrack validation set and on the partial PoseTrack test set. The individual values were obtained by submitting our predictions to the PoseTrack evaluation server. In Table 1, we show quantitative results for single-frame multi-person pose estimation with the commonly used [8, 3] mean average precision (mAP). In Table 2, we show quantitative results for multi-person pose tracking with the metrics multiple object tracking accuracy (MOTA) and multiple object tracking precision (MOTP) as used by [3].

## 6. Conclusion

We have presented a method for multi-person pose estimation and tracking. Our proposed architecture integrates multi-person detection and single-person pose estimation within a single fully convolutional network, trained in an end-to-end manner. A geometric tracker matches the predicted poses frame-by-frame to create unique IDs for each person throughout entire video sequences. Results on the PoseTrack dataset are promising and await further comparison with other methods.

## References


[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *Proc. Comput. Vis. Pattern Recognit.*, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proc. Int. Conf. Comput. Vis.*, pages 1026–1034. IEEE, 2015.

[3] U. Iqbal, A. Milan, and J. Gall. PoseTrack: Joint Multi-Person Pose Estimation and Tracking. In *Proc. Comput. Vis. Pattern Recognit.*, 2017.

[4] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *Int. Conf. Learn. Represent.*, CoRR, arXi, 2015.

[5] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proc. Comput. Vis. Pattern Recognit.*, pages 3431–3440. IEEE, 2015.

[6] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards Accurate Multi-Person Pose Estimation in the Wild. In *Proc. Comput. Vis. Pattern Recognit.*, 2017.

[7] C. Payer, D. Štern, H. Bischof, and M. Urschler. Regressing Heatmaps for Multiple Landmark Localization Using CNNs. In *Proc. Med. Image Comput. Comput. Interv.*, pages 230–238. Springer, 2016.

[8] 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 *Proc. Comput. Vis. Pattern Recognit.*, 2016.

[9] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proc. Med. Image Comput. Comput. Interv.*, pages 234–241. Springer, 2015.

[10] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In *Proc. Neural Inf. Process. Syst.*, pages 1799–1807, 2014.