Keypoint-wise Adaptive Loss for Whole-Body Human Pose Estimation

Bongjo Seong, Hyug-Jae Lee, Rokkyu Lee

AI Research Lab., NHN Cloud Corp., Seongnam-si, Gyeonggi-do, Republic of Korea
bj.seong@nhn.com, hyugjae.lee@nhn.com, rokkyu.lee@nhn.com

Abstract

This paper proposes a whole-body human pose estimation method by investigating the mixed characteristics of dense and coarse keypoints. Unlike the conventional pose estimation method, a whole-body pose estimation method needs to locate keypoints on the body as well as the face, hands, and feet. The fixed Gaussian sigma has been used in a ground-truth heatmap. Thus, whole-body pose estimation methods suffer from scale differences for each body part (i.e., different labeling noise for each body part). To address this problem, we propose a Keypoint-wise Adaptive Loss (KAL) method to learn the adaptive factors between body parts (i.e., more densely annotated face and hand keypoints than body and foot keypoints). To improve localization accuracy of dense keypoints, we further introduce Foreground-Weight Adaptive Heatmap Regression (FWAHR) method to KAL, that results in introduction of Foreground-Weight Keypoint-wise Adaptive Loss (FoWKAL). The experimental results reveal that the FoWKAL method significantly outperforms previous methods, especially on the keypoints of the body and foot, and it also achieves state-of-the-art results on COCO-WholeBody dataset.

Introduction

Human pose estimation aims to localize body keypoints in images and videos. Human pose estimation plays a critical role in visual understanding tasks and has many applications, such as human action recognition (Yan, Xiong, and Lin 2018), motion capture (Willett et al. 2020), and virtual reality (Weng, Curless, and Kemelmacher-Shlizerman 2019). Recently, beyond over-the-body pose estimation, a challening whole-body human pose estimation has been studied due to the fine-grained keypoints, complex pose, occlusion, and scale variation. Whole-body pose estimation simultaneously localizes 133 keypoints on the body as well as on the face, hands, and feet (i.e., 17 keypoints on the body, 6 on the foot, 68 on the face, and 42 on the hand). These keypoints, called COCO-WholeBody dataset (Jin et al. 2020), have different scales even for the same person (i.e., different labeling noise for each body part). For example, the hand and face have a much smaller scale than the body and foot. With these characteristics, the keypoints on the face and hands are dense, whereas the keypoints on the body and feet are coarse. In the viewpoint in heatmap regression, we observed a heatmap covering adjacent keypoints in dense keypoints as in Fig 1. It is a natural way to devise a method that considers the relationship between dense and coarse keypoints. Recent whole-body pose estimation researches have mostly focused on fine-grained keypoints (Jin et al. 2020; Zauss, Kreiss, and Alahi 2021; Zeng et al. 2022).

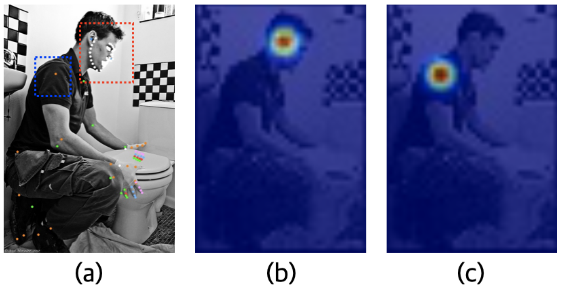


Figure 1: Comparison of a heatmap covering adjacent keypoints. (a) GT coordinates of whole-body keypoints. (b) GT heatmap covering adjacent dense keypoints. (c) Heatmap that rarely covers adjacent coarse keypoints.

We investigate a heatmap regression method focusing on keypoints with different densities. Heatmap regression methods perform better than keypoint regression methods in detecting body keypoints (Toshev and Szegedy 2014; Newell, Yang, and Deng 2016; Chen et al. 2018; Xiao, Wu, and Wei 2018; Sun et al. 2019; Cheng et al. 2020). However, different scales of whole-body keypoints have rarely been studied at the heatmap level. Simply replacing the Mean Squared Error (MSE) loss with the Adpative Wing (AWing) loss (Wang, Bo, and Fuxin 2019) is not a proper solution because we observed estimation performance degradation, except in the face and hand keypoints.

Based on this observation, we propose KAL, which learns adaptive factors to balance heatmap loss, such as MSE and AWing, according to density. The Keypoint-wise Adaptive Factor (KAF) quantifies the density of keypoints and refers to the labeling noise, i.e., annotation variance (Jin et al. 2020), inherent ambiguities (He et al. 2019; Choi et al. 2019; Luo et al. 2021). The proposed method applies more weight to foreground pixels than the background pixels if annotation variance is low while the same weight is applied to every pixel in the heatmap when annotation variance is not low.

After balancing heatmap loss between body part, however, we observed that background pixels in heatmap may aggravate the performance of dense keypoints. We introduce FWAHR method to improve localization accuracy.

The key contributions of this paper can be summarized as follows:

- We propose KAL to handle with the scale difference, ambiguity and labeling noise of keypoints, and this method experimentally shows the performance improvement of the coarse keypoints.
- To improve localization accuracy, we propose FWAHR to constrain loss weights to foreground area, which lead model to focus on foreground pixels. The proposed FWAHR induces significant performance enhancement for the dense keypoints over KAL.
- Through extensive experiments, it is shown that we achieve state-of-the-art performance with the proposed FoWKAL method.

Related Work

Human Pose Estimation. Conventional human pose estimation aims to localize 17 body keypoints, such as in the COCO keypoint challenge (Lin et al. 2014). In pose estimation, top-down and bottom-up approaches have primarily been studied (Dang et al. 2019; Chen, Tian, and He 2020). The bottom-up approach detects all keypoints simultaneously in the input image and groups the keypoints by person (Newell, Huang, and Deng 2017; Cao et al. 2017; Kocabas, Karagoz, and Akbas 2018; Cheng et al. 2020; Luo et al. 2021). Thus, it is faster than the top-down approach for multi-person pose estimation in general.

The bottom-up approach suffers from the various scales of people, whereas the top-down method is more effective and has higher accuracy for single-person pose estimation (Newell, Yang, and Deng 2016; Chen et al. 2018; Xiao, Wu, and Wei 2018; Sun et al. 2019; Huang et al. 2020). First, a human detector detects all human bounding boxes in the input image (Girshick 2015; Cai and Vasconcelos 2018; Chen et al. 2019; Tian et al. 2019). Then, a pose estimator detects 17 body keypoints of a person in a bounding box. It crops and scales for each normalized single-person pose; thus, the pose estimator has limitations on latency-constrained real-world systems, but it does not suffer from the scale variety of people.

Whole-body Human Pose Estimation. Recently known as a challenging task, whole-body dataset (Jin et al. 2020) have extended COCO dataset (Lin et al. 2014). This dataset makes whole-body pose estimation difficult due to scale variation, complex poses, mixed fine-/coarse-grained keypoints, and occlusion, requiring higher localization accuracy. Recent whole-body pose estimation researches have mostly focused on fine-grained keypoints. As a top-down approach, ZoomNet (Jin et al. 2020) is a single network consisting of multiple branches to zoom-in on fine-grained keypoints. TCFormer (Zeng et al. 2022) introduces transformer-based architecture to focus on various sizes of body part rather than background in the input image. By clustering tokens, TCFormer generates not fixed but dynamic vision tokens. As a result, its capability to estimate fine-grained keypoints is improved. In a bottom-up approach, keypoint communities (Zauss, Kreiss, and Alahi 2021) introduces a skeleton-based graph to assign different weights for each body part. Unlike previous methods that assign the same weight to each keypoint, the method based on the graph community concept effectively predicts fine-grained keypoints and various poses by quantifying the connection strength of adjacent parts, outperforming previous methods.

Loss functions for Whole-body Human Pose Estimation. Convolutional neural network based heatmap regression has been widely studied for keypoint localization in human pose estimation. Unlike the keypoint coordinate (Toshev and Szegedy 2014; Carreira et al. 2016), the keypoint heatmap represents probability with being joints as two-dimensional (2D) Gaussian kernels with a fixed sigma (Tompson et al. 2014; Wei et al. 2016; Newell, Yang, and Deng 2016; Chen et al. 2018), and therefore, research has been mainly reconstructing high-resolution heatmaps (Tompson et al. 2015; Newell, Yang, and Deng 2016; Chen et al. 2018; Xiao, Wu, and Wei 2018; Sun et al. 2019). However, loss functions have rarely been studied in human pose estimation. The MSE loss has primarily been used.

We use the keypoint heatmap to focus on the two crucial problems in the previous whole-body pose estimation methods. i) The keypoint heatmap has an imbalance problem between foreground and background pixels, degrading the model performance (i.e., foreground pixels are dominated by background pixels). ii) Whole-body keypoints are mixed dense/coarse but are encoded into the heatmap as a 2D Gaussian distribution with the same sigma σ. Consequently, using only the MSE loss, which applies the same weight the loss to all pixels, creates a problem. This paper proposes KAL for whole-body human pose estimation inspired by AWing loss (Wang, Bo, and Fuxin 2019) to address the above problems. The proposed method, inspired by the loss balanced strategy in (Kendall, Gal, and Cipolla 2018), learns to balance between each body part to cope with the labeling variance and learns to focus more on the foreground pixels than the background pixels in the heatmap.

Method

This section introduces KAL, which follows the top-down pipeline (Zhang et al. 2020). As a two-stage method, it consists of a human detector and pose estimator. The proposed method is corresponding to a pose estimator. In addition, KAL is based on the HRNet (Sun et al. 2019) backbone network and Dark (Zhang et al. 2020) as the data processing method. KAL predicts all keypoints for the single-person image. Given an image I of size $W \times H \times 3$, a pose estimator predicts the heatmap P of size $W' \times H' \times K$. After postprocessing, it detects all keypoint K from a heatmap. To improve accuracy by learning with KAL, we added a head network to predict the keypoint-wise adaptive factors

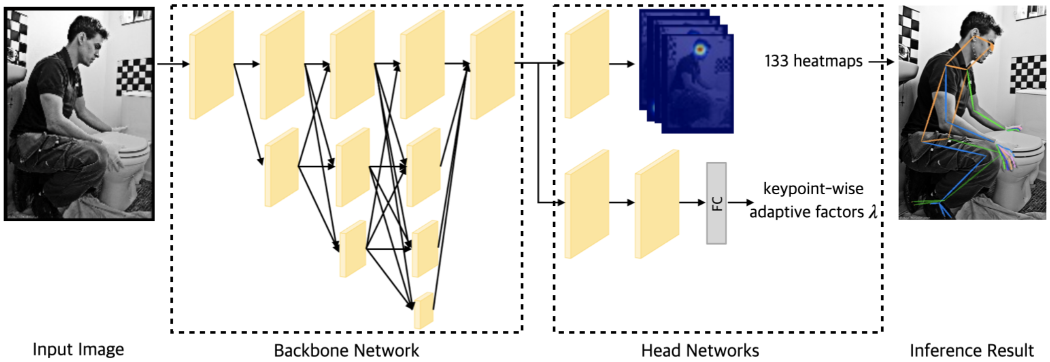


Figure 2: A human pose estimator architecture of the proposed KAL/FoWKAL method.

$\lambda \in [0, 1]^K$. The network architecture of KAL is depicted in Fig. 2.

Heatmap Regression

In heatmap regression, the ground-truth heatmap is the 2D Gaussian distribution centered at the labeled keypoint coordinate that has the fixed sigma σ. In the experiments, we used $\sigma = 2$. Suppose $P_k \in \mathbb{R}^{W' \times H'}$ denotes the ground-truth heatmap where k is the index of keypoint, $k \in \{1, ..., K\}$. $p_k^{i,j}$ denotes the ground-truth probability of the corresponding (i, j) pixel being kth keypoint, where $i \in \{1, ..., W'\}$ and $j \in \{1, ..., H'\}$. We simply denote the pixel value as p instead of $p_k^{i,j}$. Commonly, heatmap regression uses the MSE as the loss function, formulated as follows:

$$L_{MSE}(P, \hat{P}) = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{H'} \sum_{i=1}^{W'} (p_k^{i,j} - \hat{p}_k^{i,j})^2, \quad (1)$$

where $\hat{P}_k$ is the predicted heatmap of the kth keypoint.

Motivation. Now, we focus on the fundamental problem, the imbalance between foreground and background pixels in the heatmap. The MSE loss applies equal weight at the pixel level. However, the background pixels dominate the foreground pixels on the heatmap, and the prediction accuracy suffers. The proposed alternative is to consider a loss function that applies more weight on the foreground pixels than the background pixels in the heatmap to improve the accuracy of locating the keypoint, inspired by AWing loss (Wang, Bo, and Fuxin 2019). The AWing loss has larger gradient than MSE loss when the foreground error is small. In other cases, AWing loss has characteristics similar to MSE loss. The AWing loss is defined as follows:

$$AWing(p, \hat{p}) = \begin{cases} w \ln(1 + |\frac{p-\hat{p}}{\epsilon}|^{\alpha - p}) & if |p - \hat{p}| < \theta, \\ A|p - \hat{p}| - C & otherwise, \end{cases} \quad (2)$$

where p and $\hat{p}$ indicate the ground-truth and predicted pixel values in the heatmap, $w, \theta, \epsilon, \alpha$ are hyper parameter with positive values, $A = w(1/(1 + (\theta/\epsilon)^{\alpha - p}))(\alpha - p)((\theta/\epsilon)^{(\alpha - p - 1)})(1/\epsilon)$, and $C = (\theta A - w \ln(1 + (\theta/\epsilon)^{\alpha - p}))$ are responsible for making the loss function continuous at the point where $|p - \hat{p}| = \theta$. In the experiments, we followed the parameter settings used by (Wang, Bo, and Fuxin 2019). We replaced the MSE loss with the AWing loss, formulated as follows:

$$L_{AWing}(P, \hat{P}) = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{H'} \sum_{i=1}^{W'} AWing(p_k^{i,j}, \hat{p}_k^{i,j}). \quad (3)$$

However, as introduced in the experimental section, performance does not improve except in the face and hands.

Heuristic Loss. Inspired by the AWing loss results on the COCO-WholeBody dataset (Jin et al. 2020), we assume that the fine-grained body parts have the advantage of focusing on the foreground for accurate predictions, but the coarse body parts suffer from label ambiguities (He et al. 2019; Choi et al. 2019; Luo et al. 2021). Therefore, we designed a heuristic loss function, where the parts of the body and feet adopt the MSE loss, and the parts of the face and hands adopt the AWing loss. Thus, the performances are improved in most body parts. The second problem is to use the same sigma σ for all keypoints (Wei et al. 2016; Newell, Yang, and Deng 2016; Chen et al. 2018). In this case, instead of adjusting the sigma σ for each body part, the heuristic loss function indirectly adjusts the sigma σ by training the model to focus differently on the foreground or background in the heatmap. The loss function of two parts (i.e., the body and feet, the face and hands), defined as follows:

$$L_{bf}(P, \hat{P}) = \frac{1}{N_{bf}} \sum_{k \in body_{kpt} \cup foot_{kpt}} \sum_{j=1}^{H'} \sum_{i=1}^{W'} (p_k^{i,j} - \hat{p}_k^{i,j})^2,$$

$$L_{fh}(P, \hat{P}) = \frac{1}{N_{fh}} \sum_{k \in face_{kpt} \cup hand_{kpt}} \sum_{j=1}^{H'} \sum_{i=1}^{W'} AWing(p_k^{i,j}, \hat{p}_k^{i,j}), \quad (4)$$

where $N_{bf} = N_{body} + N_{foot}$, $N_{fh} = N_{face} + N_{hand}$, N_{part} denotes the number of keypoints of a body part (e.g., $N_{body} = 17$), $part \in \{body, face, foot, hand\}$ indicates each body part name, and $part_{kpt}$ denotes a set of keypoints of the corresponding body part. Now, the heuristic loss function using Eq. 4, is defined as follows:

$$L_{heuristic}(P, \hat{P}) = \lambda_{fh} L_{bf}(P, \hat{P}) + \lambda_{bf} L_{fh}(P, \hat{P}), \quad (5)$$

Method	whole-body		body	foot	face	hand
	AP	AR	AP	AP	AP	AP
Bottom-up methods:						
AE (Newell, Huang, and Deng 2017)	27.4	35.0	40.5	7.7	47.7	34.1
OpenPose (Cao et al. 2017)	33.8	44.9	56.3	53.2	48.2	19.8
Keypoint Communities (Zauss, Kreiss, and Alahi 2021)	60.4	-	69.6	63.4	85.0	52.9
Top-down methods:						
ZoomNet† (Jin et al. 2020)	54.1	65.8	**74.3**	**79.8**	62.3	40.1
HRNet-w32 (Sun et al. 2019)	55.3	62.6	70.0	56.7	63.7	47.3
TCFormer (Zeng et al. 2022)	57.2	67.8	69.1	69.8	64.9	**53.5**
HRNet-w32+DARK (Zhang et al. 2020)	58.2	67.1	69.4	56.5	73.6	50.3
HRNet-w32+DARK+FoWKAL (Ours)	**61.6**	**71.1**	72.7	74.2	**73.8**	**53.5**

Table 1: Performance comparisons with the state-of-the-art bottom-up/top-down methods. The results are reported on the COCO-WholeBody V1.0 dataset (Jin et al. 2020). HRNet-w32 and HRNet-w32+DARK results are from MMPose (Contributors 2020). ZoomNet† is trained with the COCO-WholeBody V0.5 training set.

where $\lambda_{fh} = N_{fh}/N$ and $\lambda_{bf} = N_{bf}/N$ are balancing factors because the number of keypoints for each body part is different, and $N = \sum_{part} N_{part}$. The above heuristic loss function aims to accurately estimate fine-grained keypoints by focusing on the foreground in the heatmap.

Keypoint-wise Adaptive Loss. In the previous section, we empirically chose a loss function between the MSE loss and AWing loss for each body part. To cope with variant poses, we introduce a Keypoint-wise Adaptive Factor (KAF) $\lambda \in [0,1]^K$ and control the extent of the focus on the foreground in the heatmap. We define the adaptive loss function at each keypoint as follows:

$$\begin{aligned} &L_{Adaptive}(P_k, \hat{P}_k) = \\ &\lambda_k L_{AWing}(P_k, \hat{P}_k) + (1-\lambda_k) L_{MSE}(P_k, \hat{P}_k), \end{aligned} \tag{6}$$

where λ_k indicates a KAF of the kth keypoint heatmap, which is the output of added head network. In addition, in the context of relationships between body parts, we add a regularization term of the KAF in the loss. The regularization term is defined as:

$$\begin{aligned} L_{reg}(\lambda) &= \sum_{part} \{ \frac{1}{N_{part}} \sum_{k \in part_{kpt}} (\lambda_k - \bar{\lambda}_{part})^2 \} \\ &= \sum_{part} Var(\lambda_{part}), \end{aligned} \tag{7}$$

where $Var(\lambda_{part})$ indicates the variance of λ_{part}, and $\bar{\lambda}_{part}$ indicates the mean of the KAF in a body part (i.e., $\bar{\lambda}_{part} = \frac{1}{N_{part}} \sum_{k \in part_{kpt}} \lambda_k$). We propose KAL, written as follows:

$$\begin{aligned} &L_{KAL}(P, \hat{P}) \\ &= \sum_{part} \{ \frac{1}{N_{part}} \sum_{k \in part_{kpt}} L_{Adaptive}(P_k, \hat{P}_k) \} + L_{reg}(\lambda). \end{aligned} \tag{8}$$

The effect of this factor is discussed in the experimental section.

Weighted Foreground Heatmap Loss. We have experimentally learned that background pixels in heatmap may aggravate the performance of dense keypoints; because all pixles in heatmap are assigned by equal weight. In (Wang, Bo, and Fuxin 2019), loss map mask focuses on foreground pixels, which leads model to improve localization accuracy. In (Luo et al. 2021), WAHR balances the fore-background samples.

To apply similar idea, we introduce Foreground-Weight Adaptive Heatmap Regression (FWAHR) and constrain loss weight to foreground pixels for improvement accuracy of localization. FWAHR down-weight the loss on the background pixels and the loss of easier samples on the foreground pixels; therefore lead the model to focus on relatively harder samples on the foreground pixels in the heatmap. FWAHR is defined as follows:

$$W(p, \hat{p}) = \begin{cases} p^{\gamma} \cdot |1-\hat{p}| + |\hat{p}| \cdot (1-p^{\gamma}) & if\ \hat{p} \geq 2^{-\frac{1}{\gamma}}, \\ \tau p & otherwise, \end{cases} \tag{9}$$

where $2^{-\frac{1}{\gamma}}$ is threshold of soft boundary to determine that a samples becomes a positive or negative sample and τ is the hyper-parameter that down-weight the sample that is likely to be on background pixels. We use $\tau = 0.01$, and follow the parameter settings used by (Luo et al. 2021). When KAL and FWAHR are used together, it is called Foreground-Weight Keypoint-wise Adaptive Loss (FoWKAL). The weighted and adaptive Loss is defined as follows to apply FWAHR to KAL:

$$\begin{aligned} &L_{WAdaptive}(P_k, \hat{P}_k) \\ &= \lambda_k L_{WAWing}(P_k, \hat{P}_k) + (1-\lambda_k) L_{WMSE}(P_k, \hat{P}_k) \\ &= \lambda_k \sum_{j=1}^{H'} \sum_{i=1}^{W'} W(p_k^{i,j}, \hat{p}_k^{i,j}) AWing(p_k^{i,j}, \hat{p}_k^{i,j}) \\ &+ (1-\lambda_k) \sum_{j=1}^{H'} \sum_{i=1}^{W'} W(p_k^{i,j}, \hat{p}_k^{i,j})(p_k^{i,j} - \hat{p}_k^{i,j})^2. \end{aligned} \tag{10}$$

Method	MSE	AWing	KAL	FWAHR	whole-body AP	body AP	foot AP	face AP	hand AP
(a)	√				58.2	69.4	56.5	73.6	50.3
(b)		√			57.9	67.6	52.4	**76.8**	50.9
(c)	√	√			58.7	70.2	58.6	76.5	48.4
(d)			√		58.4	71.8	73.4	69.6	45.8
(e)	√	√		√	61.2	71.1	69.0	76.4	53.2
(f)			√	√	**61.6**	**72.7**	**74.2**	73.8	**53.5**

Table 2: Ablation study on Mean Squared Error/Adaptive Wing loss, Keypoint-wise Adaptive Loss (KAL), and Foreground-Weight Adaptive Heatmap Regression (FWAHR), respectively. Method (a) is the baseline with MSE loss, method (b) is the AWing loss, method (c) is the heuristic loss, method (d) is the KAL, method (e) is the heuristic loss and the FWAHR, and method (f) is the Foreground-Weight Keypoint-wise Adaptive Loss (FoWKAL).

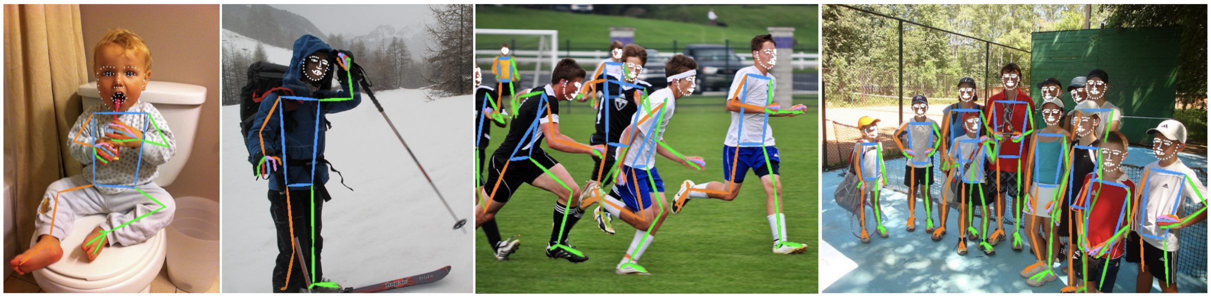


Figure 3: Qualitative results with FoWKAL from the COCO-WholeBody V1.0 validation set (Jin et al. 2020).

Finally, by modifying Eq. (8), FoWKAL is defined as follows:

$$L_{FoWKAL}(P, \hat{P}) \\ = \sum_{part} \{ \frac{1}{N_{part}} \sum_{k \in part_{kpt}} L_{WAdaptive}(P_k, \hat{P}_k) \} + L_{reg}(\lambda). \quad (11)$$

Experiments and Analysis

Datasets and Evaluation metric

We conducted extensive experiments on the COCO-WholeBody V1.0 dataset (Jin et al. 2020), including 118K training images and 5K validation images. This dataset contains annotations of 17 body, 6 foot, 68 face, and 42 hand images for the human pose. The standard evaluation metrics for whole-body human pose estimation use Average Precision (AP) and Average Recall (AR) based on object keypoint similarity (Lin et al. 2014; Jin et al. 2020).

Implementation Details

Training and Testing. The proposed method consists of HRNet-W32 (Sun et al. 2019) as the backbone network and Dark (Zhang et al. 2020) as the data processing method. We train the backbone networks initialized by the model pre-trained for the ImageNet classification (Deng et al. 2009). For training, we follow most of the default settings of training and evaluation, as in MMPose (Contributors 2020). Specifically, the input size of the network is 256×198, and the network output is a 64×48 heatmap. Data augmentation includes random rotation ([-40°, 40°]), random scale ([0.5, 1.5]), random flip, and half body augmentation (Wang et al. 2018). We used the Adam optimizer (Kingma and Ba 2014) and choose the linear warm-up strategy. The warm-up iteration is set to 500, and the warm-up ratio is 0.001. The base learning rate is 5e-4 and drops to 5e-5 and 5e-6 at the $170th$ and $200th$ epochs, respectively. The proposed method is trained on 8 GPUs with a batch size of 32 in each GPU, and the training process is terminated within 210 epochs.

For test purposes, we follow the process proposed in previous work (Sun et al. 2019; Xiao, Wu, and Wei 2018; Newell, Yang, and Deng 2016; Chen et al. 2018) to create the heatmap by averaging the heatmaps of the original and flipped images.

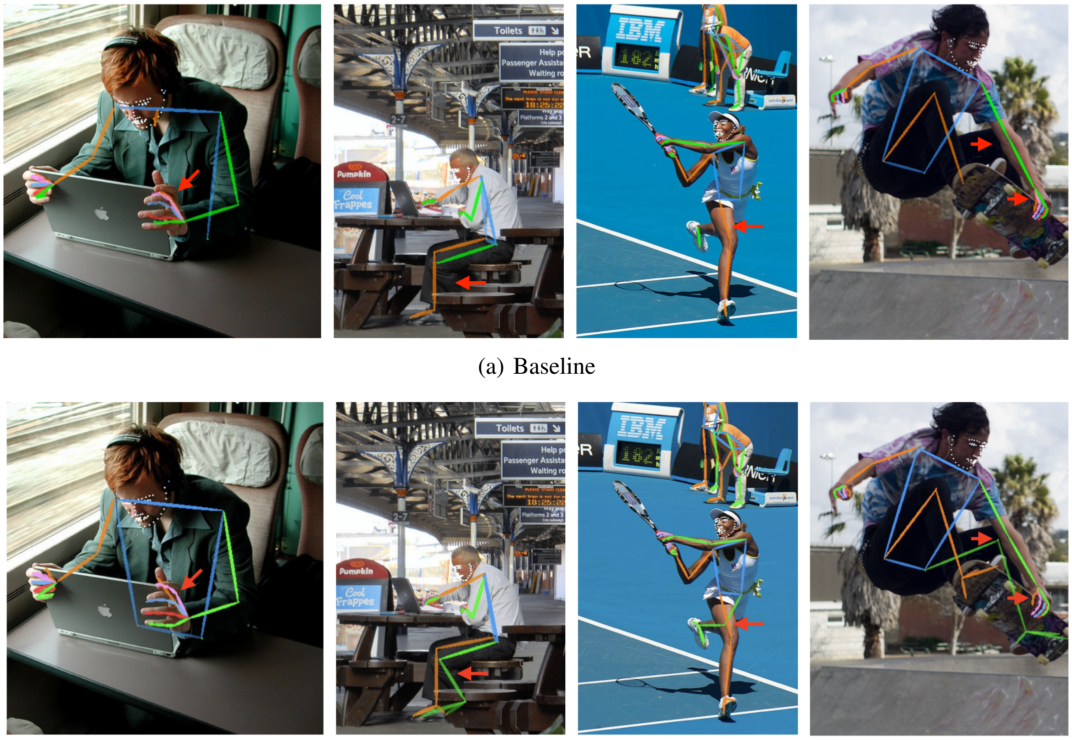
Results on the COCO-WholeBody V1.0 dataset

The proposed FoWKAL method achieves the best performance compared to previous state-of-the-art methods. The results are presented in Table 1. Compared to top-down methods like the proposed method, the FoWKAL method outperformed the previous best results (TCFormer) by 4.4% for whole-body AP (Zeng et al. 2022).

Moreover, compared to the bottom-up methods, the FoWKAL method exceeds the previous state-of-the-art results (Keypoint Communities) by 1.2% of the whole-body AP. The AP of the body and foot results significantly exceed 3.1% and 10.8%, respectively, outperforming Keypoint Communities (Zauss, Kreiss, and Alahi 2021). The qualitative results are depicted in Fig. 3.

Ablation Study

We conduct comparative experiments to validate the improvements by the MSE/AWing loss, KAL, and FWAHR.



(a) Baseline

(b) The proposed FoWKAL

Figure 4: Qualitative comparison results between FoWKAL and the baseline corresponding to Method (a) in Table 2. (a) Results on the HRNet (Sun et al. 2019) and Dark (Zhang et al. 2020) with MSE loss. (b) Results on the proposed FoWKAL method. Note the difference denoted by the red arrow.

The results are presented in Table 2 and are evaluated on COCO-WholeBody V1.0 validation set (Jin et al. 2020).

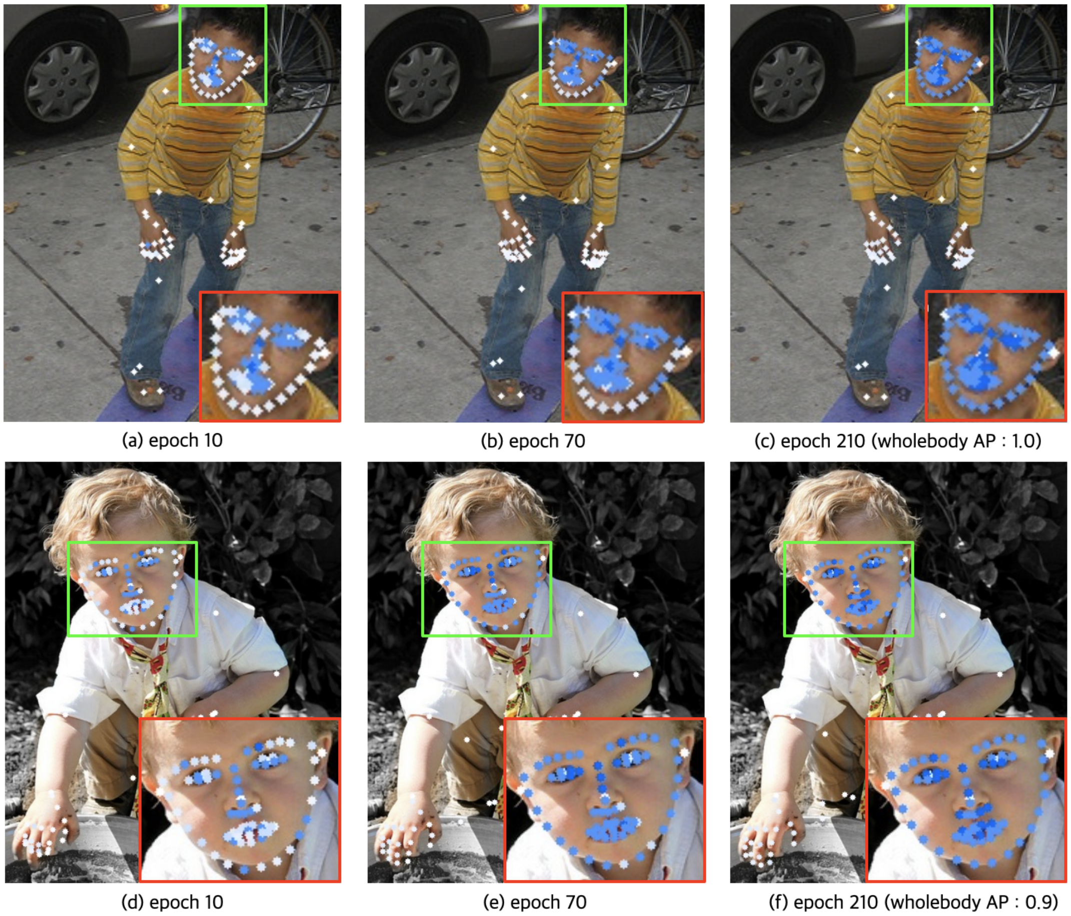
Dense/Coarse keypoints The baseline is Method (a) with the MSE loss function and equally weight the loss at the pixel level. To cope with dense keypoints, we replaced the MSE loss with AWing loss (Wang, Bo, and Fuxin 2019). In the Method (b) results, AWing loss improves AP values of face and hand. Focusing on the foreground pixel errors is effective on dense keypoints, whereas performance for coarse keypoints is degraded. Inspired by the results of Method (b), Method (c) is a heuristic loss corresponding to Eq. (5). Heuristic loss improves the performance of each body part AP and improves performance by 0.5% of the whole-body AP over the baseline. These results imply that keypoints from different body parts and scales have different labeling noise.

When the heuristic loss is replaced with the KAL, Method (d) improves performance of coarse keypoints, and especially the foot AP by 14.8%; however, dense keypoints are degraded. We suppose that KAL balances the heatmap loss between each body part, but suffers from dominant background pixels in a heatmap.

Effect of FWAHR on dense keypoints To explore the effect of FWAHR, we apply FWAHR to the heuristic loss and KAL. In both Method (e) and (f), significantly improve performance by 2.5% and 3.2% of whole-body AP, respectively, proving the effectiveness of the FWAHR. Notably, Method (f) that adds FWAHR to the KAL, significantly improves performance of dense keypoints and the results of the AP for the face and hands exceeded 4.2% and 7.7%, respectively; FoWKAL achieves state-of-the-art performance compared with previous methods. Figure 4 depicts the qualitative comparison of Methods (a) and (f), corresponding to the baseline and FoWKAL, respectively. As presented in Fig. 5, KAF have the critical role of adaptively focusing on body parts in the training process to control the extent of the focus on the foreground in the heatmap. The KAF makes the loss in the face close to the AWing loss. Otherwise, the loss is close to the MSE loss.

Method	Whole-body AP
Heuristic + WAHR	59.6
Heuristic + FWAHR	61.2
KAL + WAHR	59.5
KAL + FWAHR	61.6

Table 3: Comparison of Foreground-Weight Adaptive Heatmap Regression (FWAHR) and WAHR (Luo et al. 2021) with heuristic or Keypoint-wise Adaptive Loss (KAL).



(a) epoch 10

(b) epoch 70

(c) epoch 210 (wholebody AP : 1.0)

(d) epoch 10

(e) epoch 70

(f) epoch 210 (wholebody AP : 0.9)

Figure 5: Comparison of Keypoint-wise Adaptive Factors (KAF) that change as learning progresses on samples. Min-Max normalization relatively scales of KAF at each pose for visualization. A whiter point color is closer to MSE, and a bluer point color is closer to the AWing loss (Wang, Bo, and Fuxin 2019).

We empirically analyze the effect of the FWAHR and WAHR (Luo et al. 2021), and the results are listed in Table 3. KAL balances the heatmap loss between body part; however, by assigning equal weight to all pixels in the heatmap, localization accuracy result in an inferior performance. Therefore, we constrain the loss weight to focus on foreground pixels. We conducted comparison experiments in both the heuristic and KAL methods. In Table 2, i) Heuristic and FWAHR methods correspond to Method (e), and ii) KAL and FWAHR correspond to Method (f). In Table 3, both i) and ii), performance improved by 1.6% and 2.1%, respectively, proving the effect of the FWAHR.

Conclusion

We propose a Foreground-Weight Keypoint-wise Adaptive Loss (FoWKAL) method to estimate dense and coarse whole-body keypoints. The KAL method learns Keypoint-wise Adaptive Factors (KAF) and balances the loss to deal with the different scale of whole-body parts. As a result, KAL improves estimation performance for coarse keypoints. Furthermore, we propose a Foreground-Weight Adaptive Heatmap Regression (FWAHR) method to improve localization accuracy and FWAHR significantly improves performance of dense keypoints. Experiments show that the KAF plays a key role in heatmap regression for detecting whole-body keypoints with different labeling noise. It is shown that the proposed method achieves state-of-the-art performance on COCO-WholeBody dataset.

References

Cai, Z.; and Vasconcelos, N. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Carreira, J.; Agrawal, P.; Fragkiadaki, K.; and Malik, J. 2016. Human Pose Estimation With Iterative Error Feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; Loy, C. C.; and Lin, D. 2019. Hybrid Task Cascade for Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Y.; Tian, Y.; and He, M. 2020. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192: 102897.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded Pyramid Network for Multi-Person Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T. S.; and Zhang, L. 2020. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Choi, J.; Chun, D.; Kim, H.; and Lee, H.-J. 2019. Gaussian YOLOv3: An Accurate and Fast Object Detector Using Localization Uncertainty for Autonomous Driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Contributors, M. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

Dang, Q.; Yin, J.; Wang, B.; and Zheng, W. 2019. Deep learning based 2D human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6): 663–676.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Girshick, R. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

He, Y.; Zhu, C.; Wang, J.; Savvides, M.; and Zhang, X. 2019. Bounding Box Regression With Uncertainty for Accurate Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, J.; Zhu, Z.; Guo, F.; and Huang, G. 2020. The Devil Is in the Details: Delving Into Unbiased Data Processing for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; and Luo, P. 2020. Whole-Body Human Pose Estimation in the Wild. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 196–214. Cham: Springer International Publishing. ISBN 978-3-030-58545-7.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kocabas, M.; Karagoz, S.; and Akbas, E. 2018. MultiPoseNet: Fast Multi-Person Pose Estimation using Pose Residual Network. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Luo, Z.; Wang, Z.; Huang, Y.; Wang, L.; Tan, T.; and Zhou, E. 2021. Rethinking the Heatmap Regression for Bottom-Up Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13264–13273.

Newell, A.; Huang, Z.; and Deng, J. 2017. Associative Embedding: End-to-End Learning for Joint Detection and Grouping. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked Hourglass Networks for Human Pose Estimation. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision – ECCV 2016*, 483–499. Cham: Springer International Publishing. ISBN 978-3-319-46484-8.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; and Bregler, C. 2015. Efficient Object Localization Using Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tompson, J. J.; Jain, A.; LeCun, Y.; and Bregler, C. 2014. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Toshev, A.; and Szegedy, C. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, X.; Bo, L.; and Fuxin, L. 2019. Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wang, Z.; Li, W.; Yin, B.; Peng, Q.; Xiao, T.; Du, Y.; Li, Z.; Zhang, X.; Yu, G.; and Sun, J. 2018. Mscoco keypoints challenge 2018. In *Joint Recognition Challenge Workshop at ECCV*, volume 2018, 4.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional Pose Machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Weng, C.-Y.; Curless, B.; and Kemelmacher-Shlizerman, I. 2019. Photo Wake-Up: 3D Character Animation From a Single Photo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Willett, N. S.; Shin, H. V.; Jin, Z.; Li, W.; and Finkelstein, A. 2020. Pose2Pose: Pose Selection and Transfer for 2D Character Animation. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, 88–99. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371186.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple Baselines for Human Pose Estimation and Tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Zauss, D.; Kreiss, S.; and Alahi, A. 2021. Keypoint Communities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11057–11066.

Zeng, W.; Jin, S.; Liu, W.; Qian, C.; Luo, P.; Ouyang, W.; and Wang, X. 2022. Not All Tokens Are Equal: Human-Centric Visual Analysis via Token Clustering Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11101–11111.

Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; and Zhu, C. 2020. Distribution-Aware Coordinate Representation for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.