ISSN 2464-4617 (print)
ISSN 2464-4625 (online)
Computer Science Research Notes - CSRN
http://www.wscg.eu
WSCG 2025 Proceedings

The Influence of Faulty Labels in Data Sets on Human Pose Estimation

Arnold Schwarz
Berlin University of Applied Sciences and Technology Berlin, Germany
arnold.schwarz@bht-berlin.de

Levente Hernadi
Berlin University of Applied Sciences and Technology Berlin, Germany
leventejanko.hernadi@bht-berlin.de

Felix Biessmann
Berlin University of Applied Sciences and Technology Berlin, Germany
felix.biessmann@bht-berlin.de

Kristian Hildebrand
Berlin University of Applied Sciences and Technology Berlin, Germany
kristian.hildebrand@bht-berlin.de

Abstract

In this study we provide empirical evidence demonstrating that the quality of training data impacts model performance in Human Pose Estimation (HPE). Inaccurate labels in widely used data sets, ranging from minor errors to severe mislabeling, can negatively influence learning and distort performance metrics. We perform an in-depth analysis of popular HPE data sets to show the extent and nature of label inaccuracies. Our findings suggest that accounting for the impact of faulty labels will facilitate the development of more robust and accurate HPE models for a variety of real-world applications. We show improved performance with cleansed data.

Keywords

Human Pose Estimation, Data Sets, Data Quality

1 INTRODUCTION

Human Pose Estimation (HPE) has recently been used to analyse and make decisions at sporting events. These analyses of human pose in the centimetre range can be decisive. The uncertainties or the quality of the machine decision making are partly due to the underlying data quality and are usually not communicated to the analyst. While HPE has a long history of previous work, current model-based approaches mostly rely on two datasets for training. These data sets are labelled for 2D HPE, which forms the basis for 3D HPE in many newer approaches [1, 2, 3, 4, 5, 6] and is crucial for further applications, including those mentioned above.

While such standardized benchmarks are a fundamental prerequisite to scientific progress, recent studies have also highlighted problems with this approach. For one, with the ever increasing pace of innovation in the field of Machine Learning (ML), models achieve very good results in benchmarks quickly and the impact of model improvements become difficult to measure when predictive performance saturates [7]. This problem is aggravated by the fact that statistical significance of performance improvements in such benchmarks is often difficult to assess: The generalization error on the test set is typically reported as a point estimate, but it can exhibit variances often larger than the differences in predictive performance commonly reported at the top of benchmark leaderboards [8]. Hence improvements on benchmarks that do not account for the uncertainty in the generalization error estimates are sometimes difficult to evaluate [9]. Second, data quality issues in the training and test data, even in well curated and carefully controlled benchmark data sets [10], have been demonstrated to impact both model training and estimates of generalization performance [11]. In this study we argue that the impact of these problems on the advances of the state of the art in HPE have been underrepresented in the literature so far. We provide empirical evidence for data quality problems in erroneous annotation data of commonly used data sets and we demonstrate their impact on model training and generalization performance. These findings have implications for the interpretation of previously published results.

Therefore we underline the importance to look at the training and evaluation data to ensure reliability in benchmarks and real-world scenarios. In this work we contribute the following to existing methods and research:

1. We provide a comprehensive analysis of labeling errors in commonly used HPE benchmark data sets, developing both a taxonomy of annotation errors

and a simple yet effective heuristic to identify faulty and difficult annotations.

2. We evaluate the proposed heuristic's impact on data quality in established models leading public HPE benchmarks, discussing the effects of cleaned data and model improvements on popular HPE models.

2 RELATED WORK

2.1 Human Pose Estimation

There have been several surveys looking into the current state of HPE technology [12, 13, 14, 15]. HPE is generally divided into whether the model regresses the body keypoints directly as pixel coordinates or whether it uses a probabilistic prediction by outputting keypoint coordinates as heatmaps, which both show notable results [12]. Furthermore, there are two widely used paradigms 2D HPE falls into: the top-down- or the bottom-up approaches. The bottom-up approach initially identifies all visible body keypoints in an image, and subsequently matches these keypoints to the respective individuals. In contrast, the top-down method begins by detecting each person and their bounding boxes, followed by pinpointing the keypoints within each of these boxes.

2.2 Training and Benchmarking Data sets

HPE models are typically trained and evaluated on specific data sets, with their performance assessed using a variety of benchmark metrics tailored to each data set. The de facto standard among these data sets are COCO[1] (Common Objects in Context) [16] and MPII[2] (Max Planck Institute for Informatics) [17] due to their size (200.000 and 40.000 labeled poses for COCO and MPII respectively) and their variety in terms of image content. For performance measurement, different metrics are employed based on the data set. For instance, MPII commonly uses a variant of the Percentage of Correct Keypoints (PCK) metric, specifically PCKh@0.5. This metric measures keypoint detection accuracy and considers a keypoint correct if it is within a threshold distance (50% of the head segment length for PCKh@0.5) from the ground truth. COCO uses the Object Keypoint Similarity (OKS) score [10], a metric that considers the person's scale and keypoint visibility in the image, and evaluates the similarity between predicted and actual keypoints, allowing for some labeling noise. Since most models evaluate and train on COCO and MPII we analyze those in more detail in subsection 3.1.

[1] https://cocodataset.org
[2] http://human-pose.mpi-inf.mpg.de/

2.3 Advancements in Keypoint Detection

A notable advancement in pixel-wise regression and keypoint detection accuracy was achieved through HRNet [18], which addresses information loss challenges by implementing an architecture consisting of parallel subnetworks with varying resolutions. This makes HRNet a preferred backbone for state-of-the-art keypoint prediction models and is used throughout in this study. Other advances tackled the issue of accuracy loss introduced by image and annotation pre- and post-processing such as Unbiased Data Processing (UPD) [19] and Distribution-Aware Coordinate Representation (DARK) [20]. Recent leaderboards on the COCO and MPII test-set often mention the following high performing technologies: Polarized Self-Attention (PSA) [21], Residual Step Network (RSN)[22] and ViTPose [23]. PSA uses the idea of Self-Attention layers [24] coupled with the concept of polarized filtering. RSN introduced intra-level feature fusion through dense connections in its *Residual Step Blocks* to refine feature representation at sequential convolution levels and VitPose utilized the popular transformer architecture for the keypoint detection task. Other recent advancements include Pose as Compositional Tokens (PCT) [25], which represents poses as tokens encoding body sub-features, and LocLLM [26], a large language model architecture (LLM) that uses text descriptions and images for keypoint identification. In Table 1 we report scores for each method as they have been published in their respective work, unfortunately no scores for the COCO test-dev have been reported for LocLLM so far.

Method	AP	PCKh@0.5
VitPose+/VitPose-G	**81.1**%	**94.3**%
HRNet + UDP + PSA	79.4%	-
RSN	79.2%	93.0%
PCT	78.3%	92.5%
HRNet + DARK	76.2%	90.6%
HRNet + UDP	76.5%	-

Table 1: Reported results of state-of-the-art 2D human pose estimation methods for the OKS metric (COCO test-dev) and the PCKh@0.5 (MPII)

2.4 Data Quality in Machine Learning

Data quality has been recognized as one of the most important hyperparameters in ML model development. While the majority of models assume stationarity of both the data distribution as well as the label distribution, this assumption is violated in most real world applications. Data errors, or more generally shifts in the data distribution, do occur often and have been studied actively [27]. In the ML community this research is often referred to *covariate shift* if the shift is attributed to the input features [28] and *label shift*, if the shift is

Figure 1: An error taxonomy specifically for keypoint label errors in HPE. Left: Categorization and summary of the different types of localization errors. Right: A series of labeling errors highlighting their diversity.

associated with the target variable [29]. Some sources of noise in training data can have positive, regularizing effects [30]. This effect is leveraged in the augmentation techniques applied in computer vision to render the models invariant w.r.t. data shifts that do not change the semantic properties of an image. Other data quality problems have been demonstrated to have severe negative impact on generalization performance [31, 11, 32]. Consequently strategies to detect data quality problems in data sets are being investigated in various application domains [33]. A key challenge in this context remains automation [34]. Several approaches were proposed to detect data set shifts [35, 36, 37] or label shifts [29]. Other approaches aim at generation of realistic errors to improve model robustness [38, 39]. Our work is inspired by and complements recent findings that demonstrate how simple heuristics can be effective for removing label errors in computer vision benchmark data sets and significantly impact the generalization error estimates [11].

2.5 Label Noise in HPE

In the particular application domain investigated in this study, 2D HPE, label noise has been discussed by several authors and is typically being reported as:

- Missing keypoints for visible body parts [40].
- Localisation errors of keypoints [41, 42, 10].
- Structure confusion (Left/Right, Arms/Legs) [41].
- Randomly annotated keypoints [41].
- Missing or incomplete occlusion label [40].

Most approaches focus on making the HPE models more robust in the presence of noisy labels, rather than cleaning data. Johnson et al. [41] model keypoint localization errors produced by human workers as an isotropic Gaussian distribution of vertical and horizontal displacement. Structural errors are assumed to be uniformly distributed across the entire data set. Using a small subset of 'expert' annotated data they define a learning task to steer faulty annotations closer to the 'expert' truth.

Kato et al. [40] adapt the concept of knowledge distillation by using a teacher model to improve insufficient ground truth labels. A student network is then trained using improved annotations to increase its performance. To handle missing, shift, and duplicate noise in point data Wan et al. [42] proposed a new loss function that takes uncertainty about labels into consideration.

Complementing this prior work we propose to investigate in more detail the types of errors and their impact on HPE model training and generalization error estimates. While previous methods accept the errors in the data set, we show their frequencies and suggest a method for detecting those outliers. By data cleansing and showing the effect on training and evaluation, we conclude the problems in effectiveness of existing models trained and evaluated on these data sets, their benchmark in general and its implications on real-world scenarios. We show a more detailed error taxonomy for keypoint label errors in Figure 1.

3 METHOD

We determine and quantify the prevalence and type of errors in the data set using a systematic analysis. This involves a comprehensive examination of various data sets (subsection 3.1), leading to the development of a detailed error taxonomy (subsection 3.2). Utilizing this taxonomy, we then devise a strategy for detecting faulty labels, enabling us to identify and address errors across the entire data set (subsection 3.3-subsection 3.6).

3.1 Data Set Selection

Due to the variety of different training and benchmarking data sets, we decided to reduce our selection to the most relevant data sets currently used for HPE. We evaluated how many unique models and methods used which data set based on data from three surveys [13, 15, 12] which included 40 different methods from 2018 to 2022. In order to include publications after 2023 we also considered additional research working with online databases. The results can be seen in Figure 2. The outcome of this evaluation indicates that COCO and MPII are the most commonly used data sets for training and testing. Many published papers rely on the training data they provide and on their evaluation tasks to assess the

performance of their methods. The ground truth annotation data is often provided by a crowd of human annotators, for example the Amazon Mechanical Turk (AMT) platform[17]. As mentioned earlier by Johnson et al. [41], human crowd work on pose labelling is prone to errors. Some data sets like COCO model margins of human labelling noise into their evaluation metric [10], however severe labelling mistakes not accounted by the metric might still influence training and validation results.

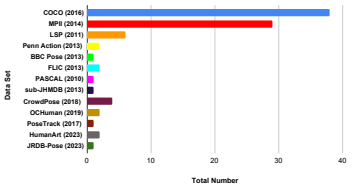

Figure 2: The plot illustrates the frequency of usage of each data set for training or benchmarking purposes. Data was gathered from three HPE surveys, encompassing 40 distinct methods, and additional research for publications after 2023.

3.2 Data Set Errors

We define an extended error taxonomy for the data sets as shown in Figure 1 and distinguish between two main error classes, localization errors and labeling errors. For localization errors, we distinguish different subclasses to describe specific types of wrong positioning of keypoints or bounding boxes. These errors are continuous, while labeling errors are discrete and describe different types of missing labels or errors in the data structure.

In order to better annotate the error frequency, we have reduced the error taxonomy (see Figure 1) to 5 classes and summarized them:

- Missing annotation: a keypoint is located on the image plane, is visible or not and is not labeled.
- False label: A keypoint is not on the image plane but is labeled.
- Incorrect position: The position of the keypoint is clearly incorrect.
- Left-right swap: Two keypoints and their assignment are swapped between left and right.
- Visibility error: The visibility flag of the key point is set incorrectly.

In a next step, we asked three people to label these five error classes in a sample of the MPII validation set (here we use the split from HRNet [18]) and the COCO data set. From the MPII validation set we randomly take 161 out of 4917 annotations and from the COCO validation set we take 163 out of 6352. This corresponds to a confidence level of 99% and a margin of error of 10%.

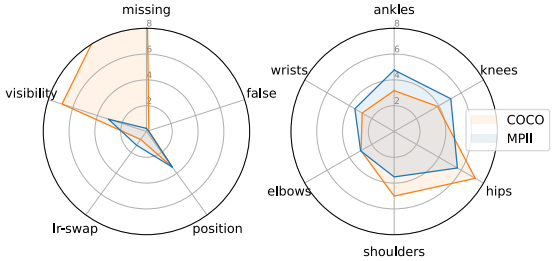

Figure 3: Left: The frequency of occurrence of an error class in our sample. Right: The frequency of error annotations for keypoint classes (only false annotations, LR swap and position errors were considered).

Our analysis reveals that both data sets contain approximately 2% positional and left-right swap errors, along with false annotations. In these cases, keypoints are incorrectly labeled, resulting in them not actually being present on the image plane (see Figure 1). As can be seen in Figure 3, both data sets show a large error for *visibility flag labels*, especially the COCO data set. Although these errors do not affect current top-down approaches, since they are not considered during training and loss calculation, they can, however, influence bottom-up approaches [43].

We should also mention that the results between the two data sets are not directly comparable, as our annotators may have been better trained to detect errors in the COCO data set, leading to a seemingly more critical approach to scoring. We did not additionally evaluate the bounding box fitting in the manual evaluation process. However, a frequency of poor bounding box fitting was noted. If we only look at the raw annotation data, we can already see that in the MPII data set around 3.4% of the annotated keypoints are outside the ground truth bounding box. For the COCO data set it is around 3.8%. If the ground truth bounding is used in the evaluation, then it may be the case that the keypoint cannot be estimated at all, as the part is cut off in the top-down process.

3.3 Automatic Label Noise Detection

In order to detect data points with faulty labels y_e automatically with high confidence we develop heuristics to estimate $p(y_e|\hat{y},y)$, the probability that a label is wrong given the prediction of a HPE model $\hat{y} \in \mathbb{R}^2$ and the (potentially faulty) annotation $y \in \mathbb{R}^2$. We assume that the majority of data points is correctly annotated and that a HPE model has learned to make accurate predictions. Large deviations between prediction $\hat{y}$ and annotations y hence indicate labeling errors. We assume that the large deviations between prediction $\hat{y}_m$ and annotations y are similar for each HPE model $m \in \{1,\dots,M\}$

described in subsection 3.4. We use the distance δ_m between the predictions $\hat{y}_m$ of model $m \in \{1,\ldots,M\}$ and the ground truth y

$$\delta_m = \hat{y}_m - y \tag{1}$$

to create a feature vector $\Delta = [\delta_1, \delta_2, \ldots, \delta_M]$. Our heuristic aggregates the distances of single joints estimates into one score by modeling the distribution of deviations $p(\Delta)$ across all models

$$p(y_e|\hat{y}, y) = 1 - p(\Delta), \tag{2}$$

For estimating $p(\Delta)$ we use a well established non-parametric outlier detection method, an *Isolation Forest* [44] as implemented in PyOD [45]. We emphasize that the choice of the outlier detection method is not the key factor for the label noise detection proposed in this work. The relevant functionality is a non-parametric density estimator that approximates the distribution of deviations $p(\Delta)$ given the multivariate feature vectors Δ. Indeed we find most other methods commonly used for outlier detection to work equally well.

3.4 Evaluation Models

We select five different top-down approach models using the MMPose[46] library to create and evaluate our label noise detection. MMPose supports various 2D human pose estimation model architectures and data sets. Therefore, it enabled us to keep evaluation standardized and fair by using the configuration files for each model listed in Table 2 and Table 3. To simplify the reproduction of results we used pre-trained model checkpoints provided by MMPose to generate the predictions for the outlier detection. The model performances on the original validation set for MPII and COCO, reported in this paper, are all consistent with scores reported by MMPose[3]. Deviations in the COCO results compared to the results listed online are due to the fact that ground truth bounding boxes were used for the COCO evaluation, to avoid an influence of the bounding box estimation errors on the metric evaluation.

We ensured models were comparable in size and performance. Model selection was mostly kept consistent between the COCO and MPII data set. However, for models pre-trained on the COCO data set there was no Hourglass model available that shared the same input size (256x192) with the other models. For COCO, we therefore, exchanged Hourglass for ResNeSt to maintain higher unity between models.

3.5 Per Keypoint Distance for MPII and COCO

Our non-parametric outlier detection requires the distance errors per joint. In order to extract this information we made the following modifications to the MPII and COCO evaluation pipelines. For MPII we modified the MMPose evaluation script so that per keypoint prediction to ground truth distance were saved in addition to the mean score values. For the COCO data set, we used a modified version of the original COCO evaluation code[4]. For each model prediction we saved the OKS metric per joint and pose as well as the distance for prediction to ground truth. Note that we did not use the 'raw' distance between prediction and ground truth but the modified version that takes the OKS σ values and the object scale into consideration. We want to note here that a person in an image can have multiple predictions, which, during the OKS calculation, are filtered using different IoU (intersection over union) thresholds. A lower threshold comes with a higher recall, which means more poses are found. For the heuristic we determined to use the 'loose' 0.5 IoU threshold to extract distance and OKS score per joint.

3.6 Calibrating the Outlier Threshold

The outlier score threshold for discarding presumably faulty annotations was calibrated as follows. We assume that model predictions are less reliable for keypoints for which there is no annotation, for instance because the body part was not on the image plane. HPE models will produce predictions for those poses, but usually these keypoints are not included in the evaluation of HPE models. Here we included these keypoints and computed outlier scores for all keypoints. In Figure 4(left) we show the distribution of all outlier scores for poses with and without keypoint annotations. In Figure 4(right) we show the outlier score distribution only for keypoints that do have an annotation. We assume that the erroneous annotations have similar characteristics to the non-annotated keypoints. Keypoints without annotations form a distinct mode of the outlier score distribution in Figure 4(left). We set the threshold for each data set individually such that outlier scores smaller than the mode of keypoints without annotations are detected as outliers. For the COCO data set the threshold was thus set to 0.75 and for MPII the threshold was set to 0.35.

4 EXPERIMENTS AND EVALUATION

We conduct a series of experiments and analysis to evaluate the effectiveness of our heuristics. We investigate how refining the validation data set influences model outcomes and assess the effects of various models on the enhanced, label-noise-free data sets. Our comparison includes examining the effect of faulty labels on evaluation metrics for COCO and MPII. We address distinguishing "hard" and "faulty" cases by contrasting manually hand-cleaned data sets (subsection 4.2) with

[3] https://mmpose.readthedocs.io/en/latest/overview.html

[4] https://github.com/cocodataset/cocoapi

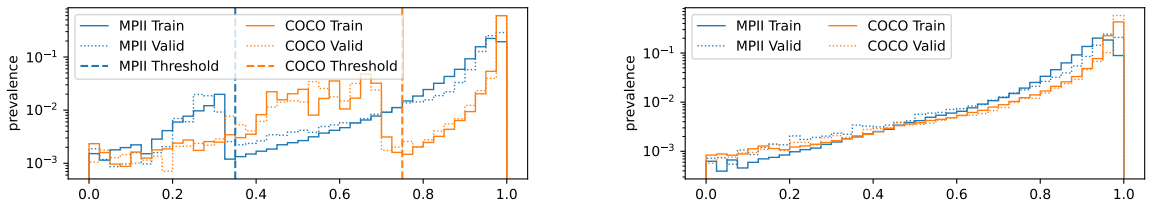

Figure 4: Determine the threshold for faulty annotations based on histograms of the outlier scores for poses on images with and without annotations. On the left side is the score distribution of all possible keypoints, including keypoints without annotations and on the right side the outlier scores only for keypoints with annotations. The inclusion of keypoints without annotations, for which a lower reliability of predictions is assumed, shows a clear second mode in the outlier score distribution (cf. left vs. right). We set the threshold for faulty annotations to exclude all annotations that (according to our heuristic) are as faulty as keypoint predictions without annotations.

sets created by our heuristics. Additionally, we examine the influence of annotation errors on the heatmaps (subsection 4.5).

4.1 Label Error Detection Evaluation

Heuristics evaluation on training set:

To evaluate the performance of the heuristic on the training data sets, we instructed two annotators to label 100 of the 4416 poses in the COCO training set and 100 of the 1102 poses in the MPII training data that were detected by our heuristic. For the COCO data our heuristic detects faulty labels with an average precision of about 36% and an average recognition rate of 31%. For the MPII data set, the heuristic detects faulty annotations with an average precision of about 41% and an average recognition rate of 25%.

Heuristics evaluation on validation set:

In order to evaluate the quality of our heuristic to detect faulty annotations in the validation data we instructed three annotators to identify faulty annotations manually on the validation data flagged as faulty by our heuristic. On validation data our heuristic detects faulty annotations with an average precision of about 16% and an average recall of 22% for the MPII validation data set and an average precision of about 15% and an average recall of 33% for the COCO validation data set.

We assume that there are several reasons why our heuristics' performance differs between the training set and the validation set. The models have much better results on the training set, than on the validation set, so separating between faulty and non-faulty keypoints should become easier. Also the annotators have different opinions on keypoint errors, concentration levels and set their focus on different keypoints/body parts. Moreover, the labeling task becomes tedious. Bazarevsky et al. [47] show that when they let two annotators relabel a data set, they achieve an average PCK@0.2 of 97.2. This raises questions on the annotation performance in HPE tasks.

4.2 Impact of Cleaning Validation Data

The results in the previous section demonstrate that the proposed heuristic reliably detects faulty labels. To evaluate the impact of erroneous labels on the HPE evaluation metrics in commonly used benchmarks, we first investigate the impact of cleaning the validation data sets. We list the results for the COCO data in Table 2 and the results for the MPII data in Table 3. For all metrics we compare results on the original validation data (RAW), the validation data cleaned automatically with our proposed heuristic, here referred to as *auto cleaned* and the impact of a partial cleanup by a human annotator (HC). We used the annotations from the error frequency analysis (see subsection 3.2) to clean the data for the HC set.

	AP			AR		
	RAW	HC	AC	RAW	HC	AC
ResNet50 [48]	73.6	73.6	74.5	76.6	76.7	77.4
ResNeSt [49]	73.8	73.8	74.7	76.8	76.8	77.6
SE-ResNet50 [50]	74.5	74.7	75.5	77.7	77.8	77.5
SCNet50 [51]	74.6	74.6	75.6	77.7	77.7	78.5
HRNet_w32 [18]	76.6	76.7	77.5	79.3	79.4	80.2
HRNet_w32 [18] TC	**76.7**	**76.8**	**77.6**	**79.4**	**79.5**	**80.2**

Table 2: Impact of cleaning on HPE metrics computed on COCO data. Compared are metrics obtained on the original data set (RAW), the manually cleaned set (HC) and the automatically cleaned data set (AC). Results obtained when cleaning training data (TC) are listed in the bottom row. Cleaning validation data leads to slight improvements, too.

For the MPII validation data set, we discard 469(1.1%) for auto clean and 161(0.4%) for hand clean and for COCO auto clean 377(0.6%) and hand clean 185(0.3%) keypoint annotations.

Our results demonstrate that discarding faulty annotations from the evaluation data improves metrics across the board slightly. In some cases the improvements are substantial. For the models trained with the MPII data set, we can see (cf. additional material) a significant

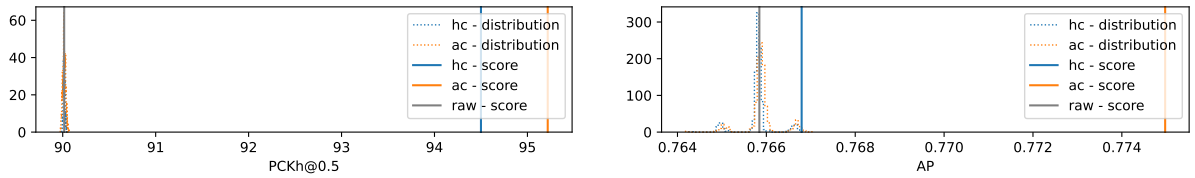

Figure 5: Comparing the accuracy impact of randomly removed keypoints across 1000 repetitions and in comparision hand-cleaned and auto-cleaned data sets. (left: PCKh@0.5 (MPII) - We achieve 509.05σ for the manually cleaned sample and 325.36σ for the automatically cleaned sample. right: mAP (COCO) - We achieve 2.97σ for the manually cleaned sample and 25.74σ for the automatically cleaned sample. The results therefore show a significant influence on the metrics.)

	PCKh@0.5			PCKh@0.1		
	RAW	HC	AC	RAW	HC	AC
ResNet50 [48]	88.2	93.6	94.4	28.6	66.7	67.3
SE-ResNet50 [49]	88.4	93.8	94.5	29.2	66.8	67.4
SCNet50 [51]	88.8	93.9	94.7	29.0	67.3	67.9
Hourglass52 [52]	88.9	94.0	94.7	31.7	68.2	68.8
HRNet_w32 [18]	90.0	94.5	95.2	33.4	**69.7**	**70.3**
HRNet_w32 [18] TC	**90.4**	**94.7**	**95.3**	**33.5**	69.6	70.2

Table 3: Impact of cleaning on HPE metrics computed on MPII data set. Compared are metrics obtained on the original validation data set (RAW), the manually cleaned set (HC) and the automatically cleaned data set (AC). Results obtained when cleaning training data (TC) are listed in the bottom rows. The input size for all models was kept consistent to 256x256

improvement in the results for the Hips, Knees and Ankles. These body parts seem to be more affected by errors in our studies (Figure 3). We can also achieve better results for the COCO data set and the benchmark metric with cleaned validation data sets. This time, however, the improvements are significantly smaller. This is especially true for the hand-adjusted validation data set. The modest improvements on COCO could be attributed to the use of a more stringent metric compared to PCKh and we cleaned a smaller subset for COCO. Nonetheless, we can also see a change in the leaderborad order for the SE-ResNet50 and SCNet50 model for the manually cleaned part.

Also, we observe a significant decrease in PCKh@0.5 score variance for the five models on the respective adjusted validation data set. For the raw data set variance was 0.38, which shrunk to 0.09 for the manually cleaned set and 0.08 for the automatically cleaned set. This suggests that the models for the MPII data set are likely to perform similarly well. No such significant decrease is observed for the COCO data set.

Influence of omitting key points on accuracy:

To measure the influence of cleaned validation data we illustrate the impact of excluding keypoints from the evaluation on metrics (PCKh and mAP) through HRNet and the corresponding validation set, as depicted in Figure 5. We randomly discarded the same number of keypoints as in the manual and automatic cleanup for the respective data sets and repeated this 1000 times. We can see a distribution of the influence of removing the random keypoints. The mean of each distribution reflects the original score (gray lines - RAW-score- in the Figure 5). We can also see that we achieve a significant score improvement with the hand-cleaned sets (HC) and the automatically cleaned sets (AC) for the MPII and COCO data set.

4.3 Impact of Cleaning Training Data

Extending previous studies [11] that only investigated the influence of cleaning on evaluation data, as we did in the previous section, we also investigated the impact of cleaning on the training data. As shown in Table 2 (COCO) and Table 3 (MPII) we observe improved predictive performance when re-training the HRNet model on cleaned data.

This effect can be seen for the model trained on cleaned MPII data for almost all body parts except the hips. In the error frequency analysis reported above, annotations for hip keypoints were often affected by labeling errors. Furthermore, hip joints have been reported before by Chen et al. [53] as "hard" to predict keypoints for models since their appearance is not always structurally obvious and often in need of more context information. We assume that similar difficulties exist for human annotators to detect hip joints correctly and therefore ground truth positions scatter a lot. Nevertheless, the overall result shows that faulty training data influences model performances.

Comparing the improvements between the two data sets COCO and MPII we find that improvements for MPII are larger. We assume that this is due to the larger training set of the COCO data set and the higher complexity in COCO.

4.4 Impact of Hard Poses

One explanation for the improvements in evaluation metrics after cleaning the data sets with the proposed

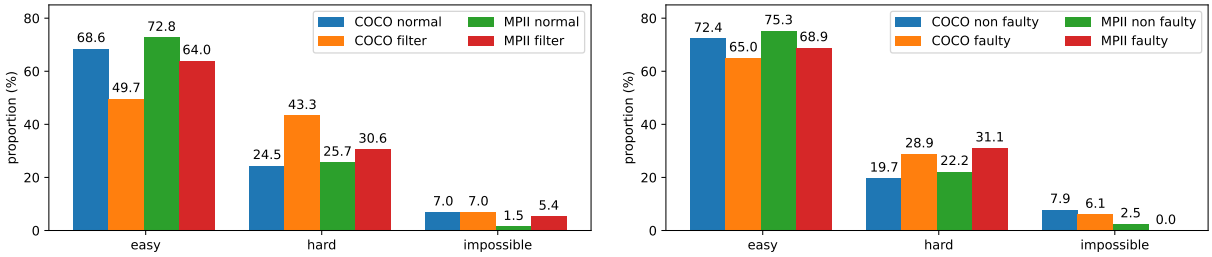

Figure 6: Evaluation of the poses of the different sets by three annotators in the categories easy, hard and impossible. (left: Comparison of the evaluation of the annotators between the sample set of the respective normal data sets and the respective sets discarded by our heuristics. right: Comparison of the evaluation of the annotators for the normal data set between manually discarded and non-discarded keypoints.)

heuristic could be that we are not (only) removing faulty annotations – but just those annotations for poses that are difficult. We investigated this hypothesis with additional experiments. Therefore, we asked three annotators to categorize images into the following three categories:

- easy: A pose is relatively easy to annotate without any further assumptions
- hard: A pose is time-consuming to annotate or assumptions have to be made
- impossible: No pose can be credibly annotated.

For the annotations of the COCO data set, we draw the ground truth bounding box on the images. For the annotations of the MPII data set, we draw the center on the image and a quadratic bounding box based on the scale information in the data. The task for the annotators is to consider only the body parts within the bounding boxes and on the image plane.

The evaluation shows that our heuristic indeed has the tendency to discard hard poses (see Figure 6(left)), but we can also see that hard poses are generally more prone to incorrect annotations, as shown in the Figure 6(right). Consequently, our heuristic tends to exclude more challenging poses, which happen to also be incorrectly annotated in many cases. Moreover, the evaluation indicates that it also omits annotations that are classified easy by human annotators, suggesting that poses easy for humans might be difficult for models. As the results of the re-training on the automatically cleaned training sets in the Table 2 and Table 3 show, both models slightly increase their performance on all relevant metrics. Even though our heuristic also partially penalizes hard poses, the effect on the models are small (0.1%). Assuming that the model does not estimate easier poses better now than before, the models do not seem to benefit from hard poses during training.

4.5 Impact of Annotation Jitter

Current approaches utilize heatmaps for the prediction of keypoints. The accuracy correlates with the variance of these heatmaps; lower variance implies higher precision, while higher variance decreases accuracy which is influenced by several factors. Our findings indicate that the variance inherent in heatmaps stems not only from the model and its hyperparameter but is also significantly influenced by the data quality. This variability is introduced by human annotators who tend to place keypoints slightly differently, an effect considered in the creation of the OKS score. To measure the impact of this annotation jitter, we employed the Human3.6M data set [54, 55], which utilizes ground truth data generated by marker-based motion capture, thereby ensuring minimal annotation jitter. We trained an HRNet model on a subset of the Human3.6M data set and introduced annotation jitter by adding a random normal distribution to the ground truth data. The perturbations were set at σ levels of 0.5%, 1%, and 2% of the bounding box diagonal.

We evaluate the 4 models based on the validation set and compress the results of the models to determine a compression ratio. We use the compression ratio as an indicator of how noisy the output heatmaps are. As shown in Figure 7, the compression ratio increases with the σ. This shows that more jitter produces noisier heatmaps. This is particularly relevant as current research [20, 56] focuses on the quality and noise of the heatmaps without investigating the reasons for this.

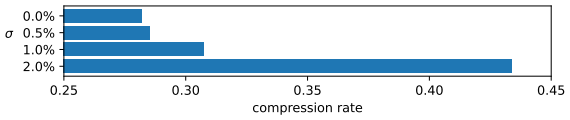

Figure 7: Comparison between annotation jitter and compression ratio using the Human3.6M dataset and HRNet. σ is the percentage of the diagonal of the bounding box used for the random normal distribution added to the ground truth data.

5 CONCLUSION

We show that the two most used data sets in the field of human pose estimation contain a variety of faulty annotations. These erroneous annotations have an impact on model training and evaluation. We are aware of the recursive chicken-and-egg problem of this work. Manual labeling by humans will always include errors. This also applies to our work. But data and research will also improve iteratively. To achieve this, we need to be aware of the errors. This should motivate further efforts to reach a better understanding of the data sets, which are commonly used in public benchmarks, and that ultimately drive scientific progress. One contribution to achieve this could be a more careful documentation of both data set creation [57] and model development [58]. Our results suggest that there is a need to improve model evaluation guidelines by creating more sophisticated testing sets, which also account for data quality in benchmark tasks. Only when we know the flaws in the data we use can we truly interpret what our models are doing and where improvements can be made. Specifically for the MPII data set and its latest leaderboard scores, we can assume that we have reached a level in this benchmark, where we can no longer achieve significantly better results. The COCO data set on the other hand, with its more robust metrics leaves more room for further improvement. Furthermore, we show that the performance of HPE models does not only depend on hard cases of poses.

ACKNOWLEDGMENTS

This research is funded by the German Research Foundation (DFG) - Project number: 528483508 - FIP 12.

REFERENCES

[1] Dario Pavllo et al. “3d human pose estimation in video with temporal convolutions and semi-supervised training”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 7753–7762.

[2] Wenkang Shan et al. “P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 461–478.

[3] Jinlu Zhang et al. “Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 13232–13242.

[4] Yu Zhan et al. “Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13116–13125.

[5] Yating Tian et al. “Recovering 3d human mesh from monocular images: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 45.12 (2023), pp. 15406–15425.

[6] Peter Hardy and Hansung Kim. “LInKs" Lifting Independent Keypoints"-Partial Pose Lifting for Occlusion Handling with Improved Accuracy in 2D-3D Human Pose Estimation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 3426–3435.

[7] Aarohi Srivastava et al. “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models”. In: *arXiv preprint arXiv:2206.04615* (2022).

[8] Xavier Bouthillier et al. “Accounting for variance in machine learning benchmarks”. In: *Proceedings of Machine Learning and Systems* 3 (2021), pp. 747–769.

[9] Peter Steinbach et al. “Machine learning state-of-the-art with uncertainties”. In: *arXiv preprint arXiv:2204.05173* (2022).

[10] Matteo Ruggero Ronchi and Pietro Perona. “Benchmarking and error diagnosis in multi-instance pose estimation”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 369–378.

[11] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. “Pervasive label errors in test sets destabilize machine learning benchmarks”. In: *arXiv preprint arXiv:2103.14749* (2021).

[12] Ce Zheng et al. “Deep learning-based human pose estimation: A survey”. In: *ACM Computing Surveys* 56.1 (2023), pp. 1–37.

[13] Luke K Topham et al. “Human body pose estimation for gait identification: A comprehensive survey of datasets and models”. In: *ACM Computing Surveys* 55.6 (2022), pp. 1–42.

[14] Pranjal Kumar, Siddhartha Chauhan, and Lalit Kumar Awasthi. “Human pose estimation using deep learning: review, methodologies, progress and future research directions”. In: *International Journal of Multimedia Information Retrieval* 11.4 (2022), pp. 489–521.

[15] Haoming Chen et al. “2D Human pose estimation: A survey”. In: *Multimedia systems* 29.5 (2023), pp. 3115–3138.

[16] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer. 2014, pp. 740–755.

[17] Mykhaylo Andriluka et al. “2d human pose estimation: New benchmark and state of the art analysis”. In: *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*. 2014, pp. 3686–3693.

[18] Ke Sun et al. “Deep high-resolution representation learning for human pose estimation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5693–5703.

[19] Junjie Huang et al. “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*. 2020, pp. 5700–5709.

[20] Feng Zhang et al. “Distribution-aware coordinate representation for human pose estimation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 7093–7102.

[21] Huajun Liu et al. “Polarized self-attention: Towards high-quality pixel-wise mapping”. In: *Neurocomputing* 506 (2022), pp. 158–167.

[22] Yuanhao Cai et al. “Learning delicate local representations for multi-person pose estimation”. In: *European conference on computer vision*. Springer. 2020, pp. 455–472.

[23] Yufei Xu et al. “Vitpose: Simple vision transformer baselines for human pose estimation”. In: *Advances in neural information processing systems* 35 (2022), pp. 38571–38584.

[24] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[25] Zigang Geng et al. “Human pose as compositional tokens”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 660–671.

[26] Dongkai Wang, Shiyu Xuan, and Shiliang Zhang. “Locllm: Exploiting generalizable human keypoint localization via large language model”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 614–623.

[27] Won Kim et al. “A taxonomy of dirty data”. In: *Data mining and knowledge discovery* 7 (2003), pp. 81–99.

[28] Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

[29] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. “Detecting and correcting for label shift with black box predictors”. In: *International conference on machine learning*. PMLR. 2018, pp. 3122–3130.

[30] Chris M Bishop. “Training with noise is equivalent to Tikhonov regularization”. In: *Neural computation* 7.1 (1995), pp. 108–116.

[31] Sebastian Schelter, Tammo Rukat, and Felix Biessmann. “JENGA: a framework to study the impact of data errors on the predictions of machine learning models”. In: (2021).

[32] Rashida Hasan and Cheehung Chu. “Noise in datasets: What are the impacts on classification performance?[noise in datasets: What are the impacts on classification performance?]” In: *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods*. 2022.

[33] Ziawasch Abedjan et al. “Detecting data errors: Where are we and what needs to be done?” In: *Proceedings of the VLDB Endowment* 9.12 (2016), pp. 993–1004.

[34] Felix Biessmann et al. “Automated data validation in machine learning systems”. In: *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* (2021).

[35] Sebastian Schelter et al. “On Challenges in Machine Learning Model Management”. In: *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* (2018), pp. 5–13.

[36] Neoklis Polyzotis et al. “Data validation for machine learning”. In: *Proceedings of machine learning and systems* 1 (2019), pp. 334–347.

[37] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. “Failing loudly: An empirical study of methods for detecting dataset shift”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[38] Görkem Algan and Ilkay Ulusoy. “Label noise types and their effects on deep learning”. In: *arXiv preprint arXiv:2003.10471* (2020).

[39] Derek Chong, Jenny Hong, and Christopher D Manning. “Detecting label errors by using pre-trained language models”. In: *arXiv preprint arXiv:2205.12702* (2022).

[40] Naoki Kato et al. “Improving multi-person pose estimation using label correction”. In: *arXiv preprint arXiv:1811.03331* (2018).

[41] Sam Johnson and Mark Everingham. "Learning effective human pose estimation from inaccurate annotation". In: *CVPR 2011*. IEEE. 2011, pp. 1465–1472.

[42] Jia Wan, Qiangqiang Wu, and Antoni B Chan. "Modeling noisy annotations for point-wise supervision". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.12 (2023), pp. 15065–15080.

[43] Yu Cheng et al. "Bottom-up 2D pose estimation via dual anatomical centers for small-scale persons". In: *Pattern Recognition* 139 (2023), p. 109403.

[44] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest". In: *2008 eighth ieee international conference on data mining*. IEEE. 2008, pp. 413–422.

[45] Yue Zhao, Zain Nasrullah, and Zheng Li. "Pyod: A python toolbox for scalable outlier detection". In: *Journal of machine learning research* 20.96 (2019), pp. 1–7.

[46] MMPose Contributors. *OpenMMLab Pose Estimation Toolbox and Benchmark*. `https://github.com/open-mmlab/mmpose`. 2020.

[47] Valentin Bazarevsky et al. "Blazepose: On-device real-time body pose tracking". In: *arXiv preprint arXiv:2006.10204* (2020).

[48] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[49] Hang Zhang et al. "Resnest: Split-attention networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 2736–2746.

[50] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[51] Jiang-Jiang Liu et al. "Improving convolutional networks with self-calibrated convolutions". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10096–10105.

[52] Alejandro Newell, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer. 2016, pp. 483–499.

[53] Yilun Chen et al. "Cascaded pyramid network for multi-person pose estimation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7103–7112.

[54] Catalin Ionescu et al. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments". In: *IEEE transactions on pattern analysis and machine intelligence* 36.7 (2013), pp. 1325–1339.

[55] Catalin Ionescu, Fuxin Li, and Cristian Sminchisescu. "Latent structured models for human pose estimation". In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 2220–2227.

[56] Yanjie Li et al. "Simcc: A simple coordinate classification perspective for human pose estimation". In: *European conference on computer vision*. Springer. 2022, pp. 89–106.

[57] Timnit Gebru et al. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[58] Margaret Mitchell et al. "Model cards for model reporting". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.