Exploring Rare Pose in Human Pose Estimation

JIHYE HWANG, JOHN YANG, AND NOJUN KWAK, (Senior Member, IEEE)
Graduate School of Convergence Science and Technology, Seoul National University, Seoul 08826, South Korea

Corresponding author: Nojun Kwak (nojunk@snu.ac.kr)

This work was supported by the IITP grant funded by the Korean Government (MSIT) under Grant 2019-0-01367, Babymind.

ABSTRACT

We tackle the issue of data imbalance between different poses in the human pose estimation problem. We explore unusual poses that are rare which occupy a small portion in a pose dataset. In order to identify a rare pose without additional learning, a simple K-means clustering algorithm is applied to a given dataset. Experimental results on MPII and COCO datasets show that outliers which are far from the nearest cluster center can be defined as rare poses and the accuracy decreases as the distance between the data point and the cluster center increases. In order to improve the performance on the rare poses, we proposed three methods for the problem of data scarcity, which are addition of rare pose duplicates, addition of synthetic rare pose data and weighted loss based on the distance from the cluster. In the proposed methods, the highest increasing score is 13.5 mAP at the rare pose data.

INDEX TERMS

Rare pose estimation, human pose estimation, deep learning, cluster distance.

I. INTRODUCTION

Due to its practical usefulness in such areas as human-computer interactions and surveillance systems, 2D pose estimation has been actively studied to efficiently locate typical human body features such as joints from images. Human pose estimation originally focused on a pose of a single person. During recent years, researches on human pose estimation tasks have been expanded to deal with multi-person poses, with which pose estimators are required to locate keypoints of multiple people that are densely visible in 2D images.

So far, many researches in the literature of 2D pose estimation have mainly proposed improved network structures or methods of efficiently using feature scale to enhance the performance of pose estimation. While the general performance of pose estimations have been gradually increased during recent years, there have been considerable differences in improvements of pose accuracy among different parts. For example, in the case of the wrist keypoint, the possible movement area of it is relatively larger due to the two stages of dependencies on an elbow and a shoulder while that of a head is much smaller due to strong correlations with two sides of shoulders. To improve the accuracy imbalance among parts, Zhu et al. applied an online hard keypoints mining loss [1]. RefineNet [2] also proposed a way to reduce the difference in accuracy among parts by learning only hard keypoints after estimating the poses on the main Network.

In this article, we extend the problem of part imbalance into a more difficult problem of pose imbalance and provide methods to deal with it. Most available pose datasets consist of data samples that are collected in daily-based situations (e.g. walking and playing sports) that are natural in motion. Huang et al. [3] had analyzed COCO [4] dataset and reported that 85% of the dataset is composed of standing poses with the rest being either sitting or lying poses.

Their work argues that the severe imbalance in the data pool makes the generalization of pose detection difficult. However, their criterion for measuring pose imbalance is mainly determined by whether a pose is stood upright. Since various factors even among the standing poses (e.g. (self-)occlusions) affect the overall pose estimation performance, a more deductive approach should be studied to quantitatively measure pose uniqueness for a better analysis of the imbalance in pose data.

In this article, we propose a method that defines rare poses for the first time and consequently propose additional techniques that lead to improvement of estimation performance against rare poses.

First, we believe an appropriate definition of special pose samples are required to solve the problem in order to enhance robustness of a pose estimator. To this end, we firstly define a rare pose as “a pose that occupies as a minority within a data population”. The examples of such rare
poses include squatting poses, poses with self-occlusion, horizontally extended poses (e.g., swimming poses) and more. A minority of a dataset, in this context, refers to outliers from the distribution of whole data.

An outlier generally means that a data sample is significantly different from others, the meaning of which is also applied for rare poses. However, unlike outliers, rare poses are not to be discarded from the set. Among various methods proposed for outlier detection, we use $K$-means clustering to detect rare poses because of its computational advantage of being a training-free clustering method. In this work, we empirically show that it is suitable to define an outlier as a rare pose which is distant from a center of clusters, unlike other data samples that are dense near the centers. Once all samples are clustered, a sample’s cluster distance (CD), the distance between a pose sample and the center point of its classified cluster, is compared with a pre-defined distance threshold (DT) value to determine whether it is a rare sample or not.

Fig 1 (a) illustrates a distribution of MPII [5] pose data samples and their clusters resulted by $K$-means clustering with $K = 7$. While the solid red arrow represents a DT,
the dashed arrow represents CD of a pose sample. If a sample’s CD is larger than DT, the pose is classified as a rare data. Fig 1 (b) shows images of rare and non-rare pose samples selected by our proposed method. Classified pose samples show clear difference of complexity between rare and non-rare poses.

Since not only the rare poses are difficult to detect, but also there exists only a scarce amount of similar data samples, we propose following three techniques to enhance the pose estimation performance:

1) Duplication of rare pose data samples. In addition to the given training data samples, we repeat rare pose data samples once more within the dataset.

2) Addition of synthetic rare pose data samples. We have created and added synthetic samples with annotations of rare pose samples to the training dataset.

3) Rarity-based loss weights. After clustering poses, the distances between poses and the center points of their corresponding clusters are used as weights for learning the amount of parameter update.

We have conducted comparison experiments among our proposed techniques to evaluate performance improvement on rare pose estimation. To further show the effectiveness of our proposed methods, we also provide quantitative results and mean average precision (mAP) scores on COCO keypoint [4] and the percentage of correct keypoints (PCKh) on MPII [5] datasets which are commonly used benchmarks of 2D multi-person pose estimation. As baselines, we have used Simple [6] and CPN [2] models which are popular networks in the multi-person pose estimation problems. From the experiments, we observed a larger increase of accuracy scores for rare pose samples.

Class imbalance means the difference in the number of data between classes in classification problems. A class that occupies a large portion of data is defined as a major class, and a class with a small number of samples is called a minor class. Over-sampling, under-sampling, weight loss and other methods are provided to solve the imbalance problem [7]–[11]. These methods put more weight on the minor class, which may reduce the performance of the major class. Therefore, it is necessary to maintain the performance of the major class. In this article, the rare pose corresponds to the minor class and our purpose is to maintain the overall performance while enhancing the performance on the rare poses.

The rest of the paper is organized as follows. We describe our definition of rare poses in Section III and propose methods to enhance performance against the rare samples in Section IV. Finally, we demonstrate the efficiency of our proposed criteria and methods for defining pose rarity and performance improvement in Section V. Conclusions are made in Section VI.

II. RELATED WORKS

A. DATA IMBALANCE

Data imbalance refers to a situation where the number of samples between classes is imbalanced. If the data is highly imbalanced, samples from the major class dominates the learning. This usually results in a model biased to the major class [9], [11].

Many researchers have proposed various methods to solve them: over-sampling, under-sampling, re-weighting the loss and synthetic minority over-sampling techniques are used [7], [8], [10], [12]–[14]. Over-sampling raises the frequency level of minor classes to the same level of major classes while under-sampling lowers the frequency of the major classes by sampling from the original distribution. Both methods can simply resolve the class imbalance problem. However, it is usually known that over-sampling and under-sampling suffers from over-fitting and under-performance problem respectively.

To overcome the over-fitting problem, synthetic minority over-sampling technique has been proposed. Instances are synthetically generated from a generator to avoid repeatedly abusing samples in the minor class. Re-weight method modifies the weight applied to each class and thus increases the importance given to the minor class. [15]–[17] Human pose datasets usually have various types of poses. Unfortunately, non-rare poses such as standing or sitting occupy a large portion of the dataset while rarely seen poses such as squatting possess smaller portion.

Therefore, we have proposed a method that could classify the type of poses with a simple criteria. After classifying poses using the proposed method, we define minor poses as ‘rare poses’. In order to improve the performance on these rare poses, we have proposed three methods inspired from methods mentioned above used to solve data imbalance problem. More detail in section IV.

B. 2D MULTI-PERSON POSE ESTIMATION

The purpose of 2D multi-person pose estimation tasks is to estimate poses of multiple people within an image. Related studies can be broadly distinguished into two ways: Top-down methods and bottom-up methods. The biggest difference between the two methods is if the estimation of the pose is performed after detection of each person. The top-down methods firstly search for human from a scene and then estimates a pose within the detected bounding box. On the other hands, bottom-up methods detect poses of multiple persons directly from input images. Due to these differences, it is generally known that the top-down methods result in higher accuracy than the bottom-up method while the single stage of bottom-up methods outperforms in computational efficiency.

As a representative research of bottom-up methods, Openpose [18] has newly proposed part affinity fields (PAFs) that express connections between body parts. The method uses the affinity fields to link multiple joints detected via heatmaps into a pose of a corresponding person in a multi-person setting. The proposed network of Openpose encodes input images into features with the VGG network structure and then projects joints to heatmaps and PAFs in a parallel manner. Part
associations are then calculated based on the maps to finally estimate multi-person poses.

The top-down methods estimate poses from objects detected by executing detection methods. Many researchers use various detection methods, of which Mask-RCNN [19] is the most commonly used one. Most previous works have done researches on utilizing multi-scale features to estimate poses for different situations and sizes. Simple [6] proposes a method to increase the scale of the output heatmaps through deconvolution layers. Upon a ResNet [20] structure, the work increases the scale of encoded features through newly added deconvolution layers. Although it is a network with a relatively simple extension, it had achieved quite good accuracy.

Many recent studies have proposed network structures that are able to utilize features in various scales concurrently. As an example, a network that exploits multi-scale features to maintain high-resolution feature scale is proposed [21]. The features in high- and low-resolution are given with separated inference paths with 4 stations to exchange information along the paths. On every last layer of each station, features are concatenated to be fed into following separated paths. For the concatenations, 1 × 1 upsampling has been applied for low-resolution features and a 3 × 3 convolution layer with a stride size of 2 has been applied to downsample high-resolution features.

Based on the analysis that local and global features are respectively important in localization and classification problems, Cai et al. [22] proposed a method that seeks to integrate both local and global features, since pose estimation problems require estimation for joint locations of different body parts. The method had achieved a state-of-the-arts performance in COCO keypoint 2017 challenge with their proposed network structure in which a convolution layer operates recursively with a single bottleneck, effectively extracting local and global features.

C. RARE POSE ESTIMATION

Localization or detection accuracy rate varies for different body parts each of which is innately given with different ranges and degrees of movement freedom. Several researches have labeled keypoints that are relatively more difficult to localize such as ankle and wrist joints. As a similar approach to the object mining method, OHEM (online hard example mining), that tries to solve data imbalance issue from object detection tasks [23], an online hard keypoints mining (OHKM) loss is proposed to solve typical accuracy imbalance among keypoints of pose estimation problems [2].

In the work, a refine network is fed with features of a global network, and both networks are trained with L2-loss functions. The refine network is applied with an OHKM loss to be trained only with parts that are detected less accurately. Another work [1] that assigns more weights on joints that are comparatively more difficult to estimate, such as partially occluded body parts, is proposed with a generative adversarial network [24] (GAN). The work collects losses for each part calculated from a generator and applies larger weights on joints with larger loss values.

As mentioned, recent papers have focused on improving pose estimation methods through utilization of the feature scales and refinement of poses using local information. While such methods have improved overall performance to an extent, a more direct approach is needed to handle poses that cover a lot of complexity. In this article, we newly define a concept of rare poses and propose methods to improve the performance of rare pose estimation tasks.

III. IDENTIFICATION OF RARE POSES

Conventionally, a pose sample that is rare represents either poses with a lot of invisible parts or an unusual pose as shown in Fig 1(b). Many methods have previously struggled from estimating such samples because of their rarity within a dataset and no clear definition to distinguish them from the usual ones. Although rare, the pose estimation for these rare poses is critical to human eyes in some areas that involve a lot of pose deformation such as gymnastics and extreme sports.

In order to improve the performance on rare poses, we firstly need to have a clear measure to identify a rare pose. Since a 2D image sample with a pose P is composed of \((x, y)\) coordinate values of \(J\) joints, i.e. \(p = \{(x_j, y_j)\}_{j=1}^{J}\), which can be considered as a 2\(J\)-dimensional continuous real random vector, it is very complicated to set a clear definition of a rare pose using the coordinate. Even if we set a heuristic rule, it takes time and cost because people have to label it. In order to solve this problem, we propose a new rare pose identification method which does not require additional learning.

The rare poses occupy a small fraction within a dataset, and have a relatively large difference from the majority of data, appearing as outliers. For the computational advantage, we aim to detect the outliers using a simple clustering method without any other additional learning of anomaly detection. In this article, we conduct the K-means clustering method [25], a popular unsupervised clustering method that searches for clusters with the minimum distance between the K cluster centers and data samples. The method allows grouping similar poses as densely as possible and labeling of rare poses which are relatively distant from the centers of clusters.

2D location information of body joints, without color and texture information, is considered to classify the poses because color and texture information tends to depend on various factors such as clothes and skin colors and thus spans an excessively wide search space. Clusters are therefore defined only by 2D coordinates \(p = \{(x_j, y_j)\}_{j=1}^{J}\) of parts from the image space which mainly represent uniqueness of each pose sample. When detecting the location information of each part, the object is positioned in the center of a certain bounding box with similar scales as illustrated in Fig. 1(a).

First, the training data are classified based on the predetermined number of clusters \(K\). In doing so, the distance between the pose \(p_i\) and the center of the corresponding cluster \(m_c\) is measured as follows which is denoted as the
cluster distance \( d_{ci} \):
\[
d_{ci} = \min_{c \in \{1, \ldots, K\}} ||p_i - m_c||
\] (1)

Then, the cluster distance is used to determine whether the pose \( p_i \) is rare pose or not. The Fig 5 (b) and (c) show the histograms of cluster distance. Both graphs confirm that the number of samples suddenly decreases from a certain value. We consider this point corresponding to a sudden drop of the number samples as the distance threshold (DT) \( \tau \) for the rare pose. We have conducted experiments to measure accuracy for this threshold setting, and more details are provided in Section V-A. Finally, the pose \( p_i \) is classified as a rare pose \( R \) or an usual pose \( U \) as follows.
\[
p_i \in \begin{cases} 
R & \text{if } d_{ci} \geq \tau \\
U & \text{otherwise} \end{cases}
\] (2)

IV. ENHANCING THE PERFORMANCE OF RARE POSE ESTIMATION

We have empirically found that the reason of low performance on rare poses is not only that they are difficult to estimate, but also that only a small amount of such samples are present in a dataset compared to relatively simpler poses. Table 1 shows accuracy of Simple [6] based on various distance thresholds between the centers of clusters and their corresponding poses. It can be seen from the results that the accuracy and the amount of rare pose data decreases as the threshold increases. With such understanding, we propose following methods to improve the performance against rare poses by focusing on the rare data: Addition of duplicates of rare pose samples and synthetic samples with rare pose labels to the training set and an objective function that reflects rarity based on the distance from the cluster centers.

**TABLE 1.** The accuracy of Simple [6] for samples of train data with cluster distance that exceeds each distance threshold (\( \tau \)).

| \( \tau \) | All (0) | <1.4 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 | 1.9 | 2 |
|---|---|---|---|---|---|---|---|---|---|
| \#data | 257252 | 249789 | 7463 | 3317 | 1395 | 551 | 217 | 95 | 45 |
| mAP | 72.4 | 80.7 | 56.5 | 37.8 | 35.4 | 30.7 | 22.8 | 15.7 |

**A. DUPLICATION OF RARE POSE SAMPLES (DRP)**

One of the effective ways to improve general performance is to train with a better-balanced dataset. To achieve a similar effect and to provide data samples from the same domain as the majority of training data, instead of collecting additional data, we have added duplicates of rare pose samples. The rare samples are firstly labeled from the training data in a preprocess and they are simply repeated once within the training set.

**B. ADDITION OF SYNTHETIC RARE POSE DATA (ASRP)**

Since data collection is expensive and collecting rare pose data is particularly more difficult, it is reasonable to synthetically generate rare samples with accompanying annotation.
ground truths if more various color/texture must be considered [27], [28].

For generations of synthetic rare pose data, we have used SMPL human body model [26]. SMPL is a mesh deformation model that is defined by pose \( \theta \) and shape \( \beta \) parameters for controlling the model’s 3D mesh outputs. The constructed 3D human mesh models from SMPL can then be projected to 2D images with camera parameters consisting of scales \( s \), translations \( t \) and rotations \( R \) to be re-created as pose data samples with 2D joint location annotations. However, since the annotations for rare pose data samples are given with image coordinates \((x, y)\) for each joint, we were required to map the 2D coordinates to the corresponding pose and camera parameters that allow creating and reprojecting SMPL mesh models in order to align the annotations of resultant synthetic samples with those of given 2D rare pose samples.

Fig 2 illustrates the overall generation process of synthetic rare pose data samples. For authenticity of human poses, SMPL provides a pool of known pose parameters and color/texture information for each mesh collected from real poses, which we utilize for generating random synthetic samples. As an initial phase, since we are required to learn a function \( f \) that maps 2D joint coordinates to corresponding SMPL parameters \( \theta, \beta, R, t, s \), we collect inputs and outputs of SMPL models in order to train \( f \) (see Fig 2(a)). With the trained \( f \) with the setting in Fig 2(b), we are able to find the right parameters that result in a 3D human mesh model with 2D annotations when reprojected to the 2D image space, as depicted in Fig 2(c). A random image then fills the background in order to create a synthetic rare pose sample which can be shown in Fig 2(d). A pool of body texture provides color values of each mesh that expresses color and wrinkle of clothes or skin. Backgrounds are randomly cropped patches from randomly selected samples of VOC2012 dataset [29]. Examples of synthetically generated pose data are shown in Fig 3. After generating the resultant synthetic pose samples \( S = \{p_{1}, p_{2}, \ldots, p_{m}\} \) are generated, the samples are added upon the given training set of ground truth poses \( P_{gt} \) so that poses that are used for training are \( P_{all} = P_{gt} \cup S \).

To generate more realistic synthesized samples, we have pre-trained a generator that translates styles from synthetic to real. We have used U-GAT-IT [30], an unsupervised generative model for image-to-image translation, for its competent performance of style transfer from cartoon to real and vice versa.

![FIGURE 3. Examples of synthetic MPII rare pose data generated by the method shown in Fig 2. Images in the first row show MPII data samples that are defined as rare poses according to our method. Images in the second row are the synthetic samples. Images in the third and forth rows show synthetically generated samples of which style is transferred from synthetic to real, respectively, with and without backgrounds.](image-url)
The model $f$ is structured as PoseResnet with ResNet50 structure from Simple [6] that takes $256 \times 256$ sized inputs. The network is selected for its reported and empirical efficiency. We have selected 13 keypoints aligning universally with SMPL, MPII and COCO datasets, so that after the network is trained with keypoints of SMPL, rare pose annotations from MPII and COCO can be used to generate corresponding synthetic samples (See Fig 3). The network is fed with 13 channels of heatmaps that are created based of 2D coordinate inputs.

C. WEIGHTED LOSS BASED ON CLUSTER DISTANCE (WLCD)

In object detection problems, soft sampling methods are applied to solve data imbalance issues [31]–[34]. The degree of contribution is assigned to a value between 0 and 1 for each data to solve data imbalance problem. Similarly, after $i$-th pose is assigned with a cluster class $c \in \{1, \ldots, K\}$ through $K$-means clustering, a cluster distance $d_i^c$, a distance between the pose and its corresponding cluster center, can be measured. The cluster distance values are applied as weights when calculating the loss, which yields larger gradient updates for rarer poses.

The weighted objective based on cluster distances of our proposed method is as follows:

$$ L = \frac{1}{N} \sum_i \sum_j (h_{ij} - \hat{h}_{ij})^2 \cdot w(d_i^c) $$

(3)

where a loss function $L$ with a weight $w(d_i^c)$ is multiplied to the mean square error (MSE) between heatmap predictions $\hat{h}_{ij}$ and ground-truths $h_{ij}$ for $j$-th joint from $i$-th pose data. Here, $N$ and $J$ are the number of training samples and the number of joints respectively. The weight is determined as follows:

$$ w(d) = \begin{cases} 1 & \text{if } d < \tau \\ 1 + (d - \tau) & \text{if } \tau \leq d < \tau + 0.5 \\ 1.5 & \text{if } d \geq \tau + 0.5. \end{cases} $$

(4)

The cluster distance is a value indicating how far the pose is from the usual pose, in other words, how the pose is rare. Even within poses classified as rare, it is possible to learn with different weights for different samples.

D. DIVIDE AND CONQUER STRATEGY FOR POSE ESTIMATION (DACP)

We have proposed $DRP$, $ASRP$ and $WLCD$ to improve the performance of pose estimation models using rare poses. The three methods have been designed for an efficient learning of the rare pose. The proposed methods generally maintain the performance of the usual pose, but some experiments have also shown results sacrificing the performance of the usual pose for the boosted performance of the rare pose, which is not a significant drop comparing to the performance gain in rare pose.

Thus, we have adopted the divide and conquer strategy to the network structure. The divide and conquer is an algorithm which recursively breaks down a problem into two or more sub-problems. To resolve the tradeoff between the performance of rare pose and usual pose at the same time, we divide our pose estimation architecture into two networks each of which focuses more on the rare pose or the others. The proposed algorithm works as below.

Algorithm 1 Divide and Conquer for Pose Estimation

1: $Net_b$: baseline network.
2: $Net_r$: baseline network with proposed methods
3: score calculates a confidence score of heatmaps.
4: postprocessing : detects the $(x, y)$ locations of pose from the heatmap.
5: for $i = 1, \ldots, N$ do
6: \hspace{1cm} $Img_i$ is an input image.
7: \hspace{1cm} Obtain $h_b$ from $Net_b(Img_i)$.
8: \hspace{1cm} Obtain $h_r$ from $Net_r(Img_i)$.
9: \hspace{1cm} if $\text{score}(h_b) < \text{score}(h_r)$ then
10: \hspace{1cm} \hspace{1cm} return postprocessing($h_r$)
11: \hspace{1cm} else
12: \hspace{1cm} \hspace{1cm} return postprocessing($h_b$)
13: \hspace{1cm} end if
14: end for

V. EXPERIMENTS

Earlier in this article, we have newly defined rare poses and proposed three strategic methods to improve the performance on the rare poses. MPII and COCO keypoints datasets are used in this section for performance evaluation of the proposed methods.

MPII and COCO keypoints are the mostly used benchmarks for training/validating 2D multi-person pose estimation models. The datasets consist of various poses from everyday poses to challenging ones.

- MPII dataset [5] has 25k images with poses of 40k people annotated with 2D locations of 16 joint parts, and they are collected based on 410 types of action categories of people. We have evaluated our method on MPII dataset with the percentage of correct keypoints [5] (PCKh) which measures a localization accuracy of the predicted joint parts. After measuring distance between the ground-truth joints and predicted joints, PCKh counts the number of joints that are within selected distance thresholds.

- COCO 2017 keypoint dataset [4] has more than 200k images with poses of 250k people which are annotated with 17 joint parts. Our methods are evaluated with mAP...
scores [4]. An object keypoint similarity [4] (OKS) is used for similarity measures among poses.

- The Leeds sport expanded dataset (LSP) [35] is a single-person pose estimation dataset. It contains dynamic sports game images such as baseball, gymnastics, tennis and so on. In this article, we have evaluated our method on test images only, numbering 1000 images.

MPII and COCO datasets are used for performance evaluation of the proposed methods, and we also test our models on LSP validation set to check the effect of the proposed methods in a different domain.

For the results of clustering, the location coordinates (x, y) of poses are normalized and used as the input feature values for clustering because the location information can classify the data regardless of the texture of the image. So, coordinates of 16 parts of MPII are used as a 32 dimensional feature vector and those of 17 parts of COCO are used as a 34 dimensional feature vector for clustering.

In order to show the effectiveness of our proposed method for performance enhancement on rare pose samples, we have set Simple [6] and CPN [2] as our baseline models. Both methods are top-down methods, and the basic structure of both methods is widely used in human pose estimation. We have conserved the network structure, hyper-parameters and the training criteria of the baseline reported except that a different batch size is used for our implementation due to our given computational resource. We use ground-truth bounding box labels of people to exclude the possibility of differences in performance caused by using an external object detector. All of the ground-truth Heatmaps are generated only using visible parts. In case of Simple [6], we adopt ResNet-50 network and input image resolution of (256,192) for COCO and (256,256) for MPII. We use data augmentations such as rescaling(±30%), rotation(±40 degrees) and flip. In case of CPN [2], we adopt the input image resolution of (256,192) for COCO and MPII. Similarly, data augmentations include rescaling(0.75~1.35), rotation(±45 degrees) and flip. In MPII and COCO, hard samples such as self-occluded poses can be frequently observed. Training the generator in ASRP, those samples participate in the training and thus the generator is able to produce challenging samples.

A. RESULTS OF RARE POSE IDENTIFICATION
Since K-means clustering is an algorithm that collects similar data by using differences among features based on K centers, its cluster classification results vary greatly depending on the number of K.

Fig 4 shows the results of experiments using different K-means clustering with different number of clusters K for.
MPII and COCO. Figs 4 (a) and (b) show the experiment results on the training sets for each dataset, and (c) and (d) show the results of experiments with each validation data on the Simple [6] baseline model. Experiments are performed by changing the number of clusters from 5 to 20 for each dataset. The x-axis represents various distance thresholds $\tau$, and the y-axis represents the resultant accuracy values for poses larger than each threshold. We also included the number of corresponding samples in the graphs. In both datasets, all clusters show a decrease in accuracy as distance threshold increases. We have chosen a relatively large number of clusters to avoid the risk of clustering to focus on a few rare poses.

We thus have selected the number of clusters to be intuitively large which also yields gradual decrements of accuracy score for a fixed threshold $\tau$ as the number of clusters increases. It is experimentally considered suitable that about 2-4% of whole dataset should be set as rare poses, which is represented as gray areas in Fig 4(a) anb (b). In the COCO case, the bar graphs in the gray section are $\tau = 1.4$ for cluster 20 / $\tau = 1.5$ for cluster 11, 15 / $\tau = 1.6$ for cluster 7 / $\tau = 1.7$ to cluster 5. Clusters 15 and 20 had low mAP with the same number of data as cluster 5, 7, and 11. This means that rare poses are not well classified as outliers when the number of clusters is too small because of the characteristics of COCO data which have many occlusions including self-occlusion. For this reason, we chose 15 clusters with $\tau = 1.5$ for COCO dataset. In the case of MPII, cluster was selected based on the same criteria. We chose 7 clusters because there were many visible parts comparing with COCO. The corresponding threshold was set $\tau = 1.0$. Finally, the values of $K$ for MPII and COCO are respectively determined as 7 and 15 through experiments.

Table 1 shows the results with various numbers for clusters. In the tables, the row ‘#data’ represents the number of samples with larger distance than a threshold $\tau$. An exception is the second column with ‘$\leq \tau$’ which tells the number of non-rare samples whose distance is smaller than the threshold $\tau$. In this context, each pose sample means one pose within a ground truth bounding box, and we had excluded COCO samples that have zero visible annotations from this experiment.

Fig 5(b) and (c) are histograms of the distance values from the cluster centers to each data respectively for MPII and COCO. The x-axis represents the percentage of the samples. For the case of MPII data of which 16 parts to be annotated if fully visible, all joints from most of its data samples (67%) are visible. Fully visible COCO data samples are defined by 17 part locations, and there are not many fully visible pose samples. Subfigures (b) and (c) represent histograms of the distance values from the cluster centers to each data respectively for MPII and COCO. In MPII graph, most data lie within cluster distances and distributed in a narrower graph width while COCO graph tends to have a larger cluster distance values and variance.

FIGURE 5. (a) shows the ratio of number of samples with certain number of visible parts out of the whole data. The x-axis represents the number of visible parts and the y-axis represents the percentage of the samples. For the case of MPII data of which 16 parts to be annotated if fully visible, all joints from most of its data samples (67%) are visible. Fully visible COCO data samples are defined by 17 part locations, and there are not many fully visible pose samples. Subfigures (b) and (c) represent histograms of the distance values from the cluster centers to each data respectively for MPII and COCO. In MPII graph, most data lie within cluster distances and distributed in a narrower graph width while COCO graph tends to have a larger cluster distance values and variance.

In the tables, the row ‘#data’ represents the number of samples with larger distance than a threshold $\tau$. Also, $\tau$ is the second column with ‘$\leq \tau$’ which tells the number of non-rare samples whose distance is smaller than the threshold $\tau$. In this context, each pose sample means one pose within a ground truth bounding box, and we had excluded COCO samples that have zero visible annotations from this experiment.

Fig 5(b) and (c) are histograms of the distance values from the cluster centers to each data respectively for MPII and COCO. The x-axis represents the percentage of the samples. For the case of MPII data of which 16 parts to be annotated if fully visible, all joints from most of its data samples (67%) are visible. Fully visible COCO data samples are defined by 17 part locations, and there are not many fully visible pose samples. Subfigures (b) and (c) represent histograms of the distance values from the cluster centers to each data respectively for MPII and COCO. In MPII graph, most data lie within cluster distances and distributed in a narrower graph width while COCO graph tends to have a larger cluster distance values and variance.

We provide the results in Table 1 to show tendency of labeling rare poses according to certain thresholds. Each table shows the number and accuracy of train data by $\tau$. Also, Fig 1(b) shows examples of rare and none-rare poses. Images that are detected as non-rare pose can be confirmed that the object has less active movements of parts with more frontal views than the ones detected as rare poses. From these results, it was confirmed that the higher the thresholds are, the lower the accuracy is with more peculiar poses are defined. Through these experiments, we have determined a reasonable thresholds ($\tau$) 1.0 and 1.5 respectively for MPII and COCO, having a reasonable amount of data classified into rare poses with a low mAP. The ‘Simple’ [6] baseline model is used to select the number of clusters and the threshold of rare pose. In the other baseline model ‘CPN’ [2], the number of clusters and the threshold $\tau$ are set to be the same as for ‘Simple’.
TABLE 2. The accuracy for each distance threshold (τ) on the MPII validation.

(a) Results of Simple [6]

| #data | All(0) | <0.5 | 0.5 | 0.7 | 0.9 | 1 | 1.2 |
|-------|--------|------|-----|-----|-----|---|-----|
| Basline | 88.53 | 90.33 | 82.40 | 82.90 | 80.18 | 77.81 | 66.98 |
| DRP | 88.60(+0.07) | 90.24(-0.08) | 82.99(+0.59) | 83.71(+0.81) | 82.06(-1.88) | 79.27(+1.46) | 73.07(-6.09) |
| ASRP | 88.81(+0.28) | 90.62(+0.29) | 82.63(+0.23) | 83.45(+0.55) | 80.18(+0.00) | 77.81(+0.00) | 67.62(+0.64) |
| ASRPT | 88.68(+0.15) | 90.48(+0.15) | 82.52(+0.12) | 83.69(+0.79) | 80.89(-0.71) | 78.64(+0.83) | 70.83(+3.85) |
| WLCD | 88.84(+0.31) | 90.46(+0.13) | 83.33(+0.93) | 83.80(+0.90) | 81.00(+0.82) | 78.43(+0.62) | 68.91(+1.98) |
| DRP + WLCD | 88.60(+0.07) | 90.16(-0.17) | 83.29(+0.89) | 84.06(+1.16) | 81.71(+1.53) | 79.58(+1.77) | 70.83(+3.85) |
| ASRPT + WLCD | 88.54(+0.01) | 90.29(-0.03) | 82.56(+0.16) | 82.79(-0.10) | 80.71(+0.53) | 78.22(+0.41) | 69.87(+2.89) |
| DRP+ASRP+WLCD | 88.43(-0.09) | 90.26(+0.06) | 82.17(-0.23) | 83.66(+0.76) | 82.23(+2.05) | 80.10(+2.29) | 74.67(+7.69) |
| DACP | 88.69(+0.16) | 90.47(+0.14) | 82.60(+0.20) | 83.08(+0.18) | 80.71(+0.53) | 78.33(+0.52) | 68.91(+1.93) |

(b) Results of CPN [2]

| #data | All(0) | <0.5 | 0.5 | 0.7 | 0.9 | 1 | 1.2 |
|-------|--------|------|-----|-----|-----|---|-----|
| Basline | 85.34 | 88.52 | 74.49 | 68.36 | 59.84 | 61.04 | 48.71 |
| DRP | 85.59(+0.25) | 88.74(+0.22) | 74.86(+0.37) | 68.65(+0.29) | 60.25(+0.41) | 62.70(+1.66) | 50.96(+2.25) |
| ASRP | 85.67(+0.33) | 88.76(+0.24) | 75.14(+0.65) | 68.73(+0.37) | 60.31(+0.47) | 61.14(+0.10) | 48.07(-0.64) |
| ASRPT | 86.13(+0.79) | 89.12(+0.60) | 75.92(+1.43) | 69.63(+1.27) | 61.01(+1.17) | 61.35(+0.31) | 48.39(-0.32) |
| WLCD | 85.79(+0.45) | 89.00(+0.48) | 74.87(+0.38) | 69.69(+1.33) | 62.48(+2.64) | 62.29(+1.25) | 51.92(+3.21) |
| DRP + WLCD | 86.21(+0.87) | 89.30(+0.78) | 75.68(+1.19) | 70.73(+2.37) | 63.30(+3.46) | 64.16(+3.12) | 51.92(+3.21) |
| ASRPT + WLCD | 85.10(-0.24) | 88.92(+0.40) | 75.35(+0.86) | 69.89(+1.53) | 61.43(+1.59) | 62.48(+1.44) | 46.47(-2.24) |
| DRP+ASRP+WLCD | 85.15(-0.19) | 88.17(-0.35) | 74.89(+0.40) | 68.13(-0.23) | 59.20(-0.64) | 61.25(+0.21) | 49.35(+0.64) |
| DACP | 85.7(+0.36) | 88.70(+0.18) | 75.48(+0.99) | 68.76(+0.40) | 59.61(-0.23) | 61.14(+0.10) | 49.00(+0.29) |

B. RESULTS OF PROPOSED METHODS

The proposed methods are divided into methods with and without additional data. The methods of adding data (duplication of rare poses and addition of synthetic rare poses) are labeled as DRP and ASRP, respectively. For the DRP case, MPII has 644 poses and COCO has 3317 poses repeated within the training set. While ASRP is a way to add a newly generated synthetic image, for a fair comparison against other proposing methods, ASRP method creates and adds the same number of rare poses as DRP's added samples. During the process of ASRP, we can obtain re-calibrated pose annotations from the SMPL model. Based on the given annotations, we can calculate the bounding box coordinates and so on. ASRPT represents the method of ASRP with samples that are transferred from synthetic to real. Lastly, the method that does not alter the training set (weighted loss based on cluster distance) is referred as WLCD.

Tables 2, and 3 show the comparison results of the baseline models and our proposed methods on MPII and COCO datasets. The values in the table represents accuracy, and the value in brackets means the difference from the performance of the baseline model. At the MPII results in Table 2, the overall results mostly increases as the highest as 0.79. τ = 1.0 assigned to rare pose in gray background, all the proposed methods show increases in performance. Especially, at τ = 1.2 where cluster distance is relatively very high, the largest increment is 6.09. We also show an increasing tendency with τ < 0.5, which only covers usual poses, indicating that the proposed method does not get hindered from learning usual poses. The performance of ASRPT is higher than that of ASRP in estimating rare poses, which indicates that matching the style (real) with the training data helps improve performance. It is possible to further improve the rare pose performance when experimenting with improving the transfer performance in future research. Unfortunately, the methods of adding synthetic rare pose data showed poor performance as shown in Table 2 (b) at τ = 1.2. However, in COCO data, when the synthetic was added, the performance was improved in Table 3 (b) τ = 1.5 and 1.8, and even when the baseline network was the Simple [6] model, the performance...
was increased. It can be expected that adding the synthetic data is not a problem and the proposed methods must be adapted to the network model and data.

Table 3 shows the results of the experiments with COCO 2017 validation set. Compared to total mAP, the proposed methods increased by about 0.1–0.3 over the baseline method except for ASRP, where some decrease of performance is observed. The $\tau = 1.5$ assigned to rare pose in gray background. At the rare pose, all of the values were increased except for two methods. Furthermore, the all cases of upper $\tau = 1.6$ tend to generally increase the performances of the proposed method. Especially, the highest accuracy improvement is 13.5. Unfortunately, several methods where $\tau < 1.1$ tends to have performance diminution, but the difference is 0.1 which is not large.

**DRP** and **ASRP** are methods of data augmentation, and **WLCD** is a method to give weight to loss. It is more effect to use the method of increasing data and weight loss at the same time to improve the rare pose. Experiments were performed on the combination of **DRP + WLCD** and **ASRP + WLCD** from COCO and MPII data. **ASRP + WLCD** showed lower results than **DRP + WLCD** combination, but **DRP + WLCD** combination outperformed the method used alone. Especially, the **DRP + WLCD** in Table 2 (b) showed the highest performance for all $\tau$ when compared with others. Combining all the proposed method (**DRP + ASRP + WLCD**), we have improved the performance on rare pose for both MPII and COCO datasets compared to both baselines (CPN and Simple).

**DACP** means the result of an experiment applying the divide and conquer method. **DACP** shows meaningful improvement in both rare pose and usual pose under all experimental settings. In rare pose, the accuracy of **DACP** is less than **DRP + ASRP + WLCD**, but still higher than baseline. **DACP** generally shows improvement in any pose.

We have proposed methods of defining a rare pose and improving performance for rare pose. In some methods, there has been a slight performance drop in usual pose due to the trade-off between usual pose and rare pose. Though it is a tolerable amount of degradation, we can still resolve this issue with **DACP**, sacrificing the inference time.

### Table 3. The accuracy for each distance threshold ($\tau$) on the COCO 2017 validation.

|        | All(0) | <1.1 | 1.1 | 1.3 | 1.5 | 1.6 | 1.7 | 1.8 |
|--------|--------|------|-----|-----|-----|-----|-----|-----|
| **data** |        |      |     |     |     |     |     |     |
| Baseline | 72.4 | 76.6 | 56.3 | 50.2 | 35.9 | 31.4 | 32.2 | 31.4 |
| **DRP** | 72.6(+0.2) | 76.5(-0.1) | 56.8(+0.5) | 50.0(-0.2) | 36.3(+0.4) | 34.3(+2.9) | 33.7(+1.5) | 35.1(+3.7) |
| **ASRP** | 71.9(-0.5) | 76.2(-0.4) | 55.7(-0.6) | 49.2(-1.0) | 35.7(-0.2) | 32.7(+1.3) | 35.7(+3.5) | 39.8(+8.4) |
| **WLCD** | 72.7(+0.3) | 76.5(-0.1) | 57.0(+0.7) | 50.9(+0.7) | 35.3(-0.6) | 31.9(+0.5) | 32.2(+0.0) | 34.7(+3.3) |
| **DRP + WLCD** | 72.8(+0.4) | 76.8(+0.2) | 57.1(+0.8) | 50.7(+0.5) | 37.5(+1.6) | 33.5(+2.1) | 33.9(+1.7) | 37.7(+6.3) |
| **ASRP + WLCD** | 72.6(+0.2) | 76.5(-0.1) | 56.8(+0.5) | 50.2(+0.0) | 36.2(+0.3) | 32.4(+1.0) | 34.8(+2.6) | 40.1(+8.7) |
| **DRP+ASRP+WLCD** | 72.6(+0.2) | 76.8(+0.2) | 56.6(+0.3) | 50.0(-0.2) | 37.1(+1.2) | 34.1(+2.7) | 37.1(+4.9) | 39.7(+8.3) |
| **DACP** | 72.8(+0.4) | 77.1(+0.5) | 56.6(+0.3) | 50.3(+0.1) | 36.8(+0.9) | 34.0(+2.6) | 37.1(+4.9) | 38.1(+6.7) |

### Table 3. (b) Results with CPN [2]

|        | All(0) | <1.1 | 1.1 | 1.3 | 1.5 | 1.6 | 1.7 | 1.8 |
|--------|--------|------|-----|-----|-----|-----|-----|-----|
| **#data** |        |      |     |     |     |     |     |     |
| Baseline | 71.2 | 75.6 | 54.8 | 48.2 | 32.8 | 28.4 | 29.0 | 26.2 |
| **DRP** | 71.2(+0.0) | 75.7(+0.1) | 54.7(-0.1) | 48.9(-0.7) | 34.2(+1.4) | 31.2(+2.8) | 29.3(+0.3) | 32.6(+6.4) |
| **ASRP** | 71.1(-0.1) | 75.5(-0.1) | 54.9(+0.1) | 48.1(-0.1) | 34.9(+2.1) | 31.7(+3.3) | 33.3(+4.3) | 36.1(+9.9) |
| **WLCD** | 71.3(+0.1) | 75.6(+0.0) | 54.9(+0.1) | 48.5(+0.3) | 35.2(+2.4) | 31.4(+3.0) | 30.3(+1.3) | 32.9(+6.7) |
| **DRP + WLCD** | 71.0(-0.2) | 75.3(-0.3) | 54.8(+0.0) | 48.1(-0.1) | 35.9(+3.1) | 32.5(+4.1) | 33.0(+4.0) | 35.1(+8.9) |
| **ASRP + WLCD** | 71.2(+0.0) | 75.8(+0.2) | 54.4(-0.4) | 47.6(-0.6) | 34.7(+1.9) | 31.9(+3.5) | 33.7(+4.7) | 35.4(+9.2) |
| **DRP+ASRP+WLCD** | 71.1(-0.1) | 75.5(-0.1) | 55.0(+0.2) | 48.9(+0.7) | 35.4(+2.6) | 31.5(+3.1) | 35.5(+6.5) | 39.7(+13.5) |
| **DACP** | 71.4(+0.2) | 75.8(+0.2) | 55.3(+0.5) | 48.9(+0.7) | 34.4(+1.6) | 30.2(+1.8) | 33.9(+4.9) | 34.5(+8.3) |
Fig 6 shows results of our methods and a baseline model [6] on one of the samples of MPII validation and COCO 2017 validation. While the baseline model struggles estimating 2D pose as Fig 6 shows, our proposing techniques better performs against the rare pose sample.

Table 4 is the results of experiments on the Leeds sports pose dataset (LSP) test. The evaluation was performed using the trained Simple model [6] on COCO keypoint data without learning with LSP. In \( n_{\text{data}} \), All (1000 poses) means all of the validation data, and selected (44 poses) means the dataset.
that we chose rare pose in the validation data. The results were measured by Percentage of Correct Keypoint (PCK). In the case of Head, because COCO annotation were different with LSP, they were excluded from the comparison. All PCK increased with the average PCK (Mean) except ASRP of selected data. Among the proposed methods, WLCD showed an increasing trend from all parts. This is because the other two methods were data augmentation with the existing data domain, so there is a domain specific point. So, WLCD is more robust to domain transfer.

VI. CONCLUSION

In this article, we have proposed a new criterion for defining rare pose samples and methods to improve performance of pose estimation for the samples. The rare pose means that the pose in the data is unique and occupy a small portion within a dataset. In other words, it is an outlier in the distribution of whole data. We have applied K-means clustering to classify the outliers. By experimentally determined distance thresholds, we are able to define and classify rare poses.

We have confirmed through experiments that the accuracy on rare pose is comparably lower than other majority poses. This is because rare poses are not only difficult poses to estimate but also occupy a small portion in the whole dataset. So, we have proposed three methods to solve this issue. The first method is to duplicate the samples defined as rare pose within the training data so that they are repeated once more. The second method is to learn a model that generates synthetic rare pose samples and provide them to the training set. Finally, we have proposed a novel loss function that applies weights based on cluster distances, the distances of between a pose and its corresponding cluster center.

We evaluated the proposed method on the COCO and MPII dataset. The proposed method increases the performance of poses defined by rare pose. While we have confirmed that the performance on the rare poses achieved by the proposed methods are significantly improved, unfortunately the overall pose performance is not due to a small percentage of the rare pose data. However, it may become more effective if the ratio of rare poses is increased in order to significantly improve the overall performance. We believe that the qualitative data generated by synthetic data must be further developed and utilized for more performance improvements.

REFERENCES

[1] A. Zhu, S. Zhang, Y. Huang, F. Hu, R. Cui, and G. Hua, “Exploring hard joints mining via hourglass-based generative adversarial network for human pose estimation,” AIP Adv., vol. 9, no. 3, Mar. 2019, Art. no. 035321.

[2] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, “Cascaded pyramid network for multi-person pose estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7103–7112.

[3] Y. Huang, B. Sun, H. Kan, J. Zhuang, and Z. Qin, “Followmeup sports: New benchmark for 2D human keypoint recognition,” in Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (FCVC), Springer, 2019, pp. 110–121.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in Proc. Eur. Conf. Comput. Vis. Springer, 2014, pp. 740–755.

[5] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2D human pose estimation: New benchmark and state of the art analysis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2014, pp. 3686–3693.

[6] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 466–481.

[7] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks.” Neural Netw., vol. 106, pp. 249–259, Oct. 2018.

[8] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, “Training deep neural networks on imbalanced data sets,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2016, pp. 4368–4374.

[9] J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” J. Big Data, vol. 6, no. 1, p. 1, 2018.

[10] S. Pouyanfar, Y. Tao, A. Mohan, H. Tian, A. S. Kaseh, K. Gauen, R. Dailey, S. Aghajanazadeh, Y.-H. Lu, S.-C. Chen, and M.-L. Shyu, “Dynamic sampling in convolutional neural networks for imbalanced data classification,” in Proc. IEEE Conf. Multimedia Inf. Process. Res. (MIPR), Apr. 2018, pp. 112–117.

[11] R. Longadge and P. Gohil, “Class imbalance in data mining review,” 2013, arXiv:1305.1707. [Online]. Available: http://arxiv.org/abs/1305.1707

[12] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, “Experimental perspectives on learning from imbalanced data,” in Proc. 24th Int. Conf. Mach. Learn., 2007, pp. 935–942.

[13] C. X. Ling and C. Li, “Data mining for direct marketing: Problems and solutions,” in Proc. KDD, vol. 98, 1998, pp. 73–79.

[14] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, “SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary,” J. Artif. Intell. Res., vol. 61, pp. 863–905, Apr. 2018.

[15] K. Cao, C. Wei, A. Gaidon, N. Areshiga, and T. Ma, “Learning imbalanced datasets with label-distribution-aware margin loss,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 1567–1578.

TABLE 4. The comparison accuracy on Leeds Sports Pose validation dataset (LSP) using Simple [6] model trained on COCO.

| Methods | Parts | Ankle | Knee | Hip | Wrist | Elbow | Shoulder | Head | Mean |
|---------|-------|-------|------|-----|-------|-------|-----------|------|------|
| All     | Baseline | 91.0  | 92.0 | 90.6 | 83.5  | 86.9  | 92.3      | 23.2 | 79.9 |
|         | DRP    | 91.1(0.1) | 92.5(+0.5) | 91.0(+0.4) | 83.2(-0.3) | 87.5(+0.6) | 93.3(+1.0) | 24.0(+0.8) | 80.4(+0.5) |
|         | ASRP   | 91.5(+0.5) | 92.6(+0.6) | 90.6(+0.0) | 82.9(-0.6) | 86.7(-0.2) | 92.2(-0.1) | 25.0(+1.8) | 80.2(+0.3) |
|         | WLCD   | 91.5(+0.5) | 92.0(+0.0) | 91.3(+0.7) | 83.4(-0.1) | 87.3(+0.4) | 92.6(+0.3) | 23.4(+0.2) | 80.2(+0.3) |

| Selected | Baseline | 36.4   | 34.1 | 44.3 | 40.9  | 53.4  | 56.8      | 27.3 | 41.9 |
|          | DRP    | 46.6(+10.2) | 43.2(+9.1) | 46.6(+2.3) | 51.1(+10.2) | 53.4(+0.0) | 61.4(+4.6) | 25.0(-2.3) | 46.8(+4.9) |
|          | ASRP   | 44.3(+7.9)  | 39.8(+5.7) | 38.6(-5.7) | 37.5(-3.4) | 47.7(-5.7) | 54.5(+2.3) | 28.4(+1.1) | 41.6(-0.3) |
|          | WLCD   | 44.3(+7.9)  | 39.8(+5.7) | 47.7(+3.4) | 44.3(+3.4) | 55.7(+2.3) | 59.1(+2.3) | 28.4(+1.1) | 45.6(+3.7) |
[16] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, “Class-balanced loss based on effective number of samples,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jan. 2019, pp. 9268–9277.

[17] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, “Cost-sensitive learning of deep feature representations from imbalanced data,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 8, pp. 3573–3587, Aug. 2018.

[18] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime multi-person 2D pose estimation using part affinity fields,” 2018, arXiv:1812.08008. [Online]. Available: https://github.com/ openPose.

[19] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 770–778.

[20] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 5693–5703.

[21] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 770–778.

[22] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhou, E. Zhou, X. Zhang, and J. Sun, “Learning delicate local representations for multi-person pose estimation,” 2020, arXiv:2003.04030. [Online]. Available: https://arxiv.org/abs/2003.04030.

[23] A. Shrivastava, A. Gupta, and R. Girshick, “Training region-based object detectors with online hard example mining,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 761–769.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[25] S. Lloyd, “Least squares quantization in PCM,” IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[26] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, “GANerated hands for real-time 3D hand tracking from monocular RGB,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 49–59.

[27] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2012). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[28] Y. Cao, K. Chen, C. C. Loy, and D. Lin, “Prime sample attention in object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2020, pp. 11583–11591.

[29] B. Li, Y. Liu, and X. Wang, “Gradient harmonized single-stage detector,” in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 8577–8584.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2980–2988.

[31] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, “Imbalance problems in object detection: A review,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Mar. 19, 2020, doi: 10.1109/TPAMI.2020.2981890.

[32] S. Johnson and M. Everingham, “Clustered pose and nonlinear appearance models for human pose estimation,” in Proc. BMVC, Aug. 2010, vol. 2, no. 4, p. 5.