PoseTrans: A Simple Yet Effective Pose Transformation Augmentation for Human Pose Estimation

Wentao Jiang[1,2], Sheng Jin[3,2], Wentao Liu[2,4], Chen Qian[2]
Ping Luo[3], and Si Liu[1,5]✉

[1]Institute of Artificial Intelligence, Beihang University
[2]SenseTime Research and Tetras.AI
[3]The University of Hong Kong [4]Shanghai AI Laboratory
[5]State Key Lab. of VR Technology and Systems, SCSE, Beihang University
{jiangwentao, liusi}@buaa.edu.cn js20@connect.hku.hk
{liuwentao, qianchen}@sensetime.com pluo@cs.hku.hk

Abstract. Human pose estimation aims to accurately estimate a wide variety of human poses. However, existing datasets often follow a long-tailed distribution that unusual poses only occupy a small portion, which further leads to the lack of diversity of rare poses. These issues result in the inferior generalization ability of current pose estimators. In this paper, we present a simple yet effective data augmentation method, termed Pose Transformation (PoseTrans), to alleviate the aforementioned problems. Specifically, we propose Pose Transformation Module (PTM) to create new training samples that have diverse poses and adopt a pose discriminator to ensure the plausibility of the augmented poses. Besides, we propose Pose Clustering Module (PCM) to measure the pose rarity and select the "rarest" poses to help balance the long-tailed distribution. Extensive experiments on three benchmark datasets demonstrate the effectiveness of our method, especially on rare poses. Also, our method is efficient and simple to implement, which can be easily integrated into the training pipeline of existing pose estimation models.

Keywords: Pose Estimation, Data Augmentation

1 Introduction

Human Pose Estimation (HPE) is the task of localizing human body keypoints (also referred to as joints) from an image. It serves as a fundamental technique for numerous applications, including action recognition, pedestrian tracking, and virtual/augmented reality. Recently, deep convolutional neural networks (DCNN) [41,35,34] have achieved drastic improvements on standard benchmark datasets. To fully exploit the power of DCNN, a large number of training data is indispensable for obtaining satisfactory performance in human pose estimation.

✉ : Corresponding Author.

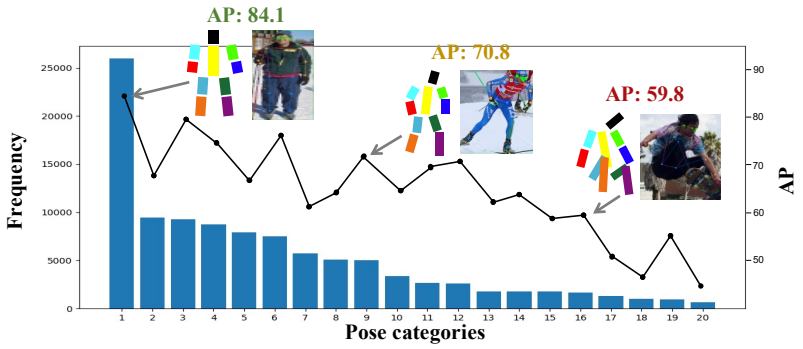

Fig. 1: We cluster the poses in the MS-COCO dataset into 20 categories and evaluate the AP with a pre-trained HRNet model [43]. The top-1 category has more than 25000 samples and high precision, while nearly half of the categories have less than 2000 samples and relatively low precision.

However, existing human pose estimation datasets do not uniformly represent all possible human poses in real life. We take MS-COCO dataset [31] as an example to analyze the distribution of the human poses, as shown in Fig. 1. We normalize the poses and cluster them into 20 categories. We observe that it follows a long-tailed distribution, with a few common pose categories (*e.g* standing and walking) occupying a large portion of the dataset and unusual posture types (*e.g* squatting and jumping) possessing a smaller portion. We also find that although current state-of-the-art data-driven methods achieve good performance on common poses, however, they still suffer performance degradation on some unusual poses, since the long-tailed categories have neither enough training samples nor enough diversity.

Due to the high cost of collecting and annotating examples with rare poses, a feasible way to tackle this problem is data augmentation. Previous methods augment the human pose mainly by global image-level transformations [37,12,35,45,42] (*e.g* scaling and rotating) or local object-level transformations [5,37,18] (*e.g* copy-paste and occluding). Since these methods fail to increase the diversity of poses and alleviate the long-tailed distribution, they contribute little to recognizing diverse rare poses.

In this paper, we propose a simple yet effective data augmentation approach, termed Pose Transformation (PoseTrans), to tackle the aforementioned challenges. PoseTrans consists of a Pose Transformation Module (PTM) with a pose discriminator, and a Pose Clustering Module (PCM). During training, PTM applies affine transformations to the original pose of the training sample and generates a pool of diverse new poses. The pre-trained pose discriminator is adopted to evaluate the plausibility of generated samples and then filter out unnatural samples. PCM is based on the Gaussian Mixture Model (GMM), which normalizes and clusters the human poses in the dataset. The rare types of poses

are represented by the Gaussian components that have small weights. PCM evaluates the components' density for each candidate pose and selects the "rarest" one (*i.e* which has the minimal weighted sum of components' density) as the final augmented training sample. By transforming the existing poses, PoseTrans helps generate diverse, plausible poses by PTM and alleviate the long-tail distribution problem by PCM. We also design a metric that focuses on rare poses called balanced AP/AR and observe more performance gain on this metric. Our method is simple to implement and can be easily integrated into the training pipeline of existing pose estimation models.

We summarize our contributions as follows:

- We present a simple yet effective data augmentation method, termed PoseTrans. To tackle the problem of limited diversity of unusual human poses, we propose a novel Pose Transformation Module (PTM) with a pose discriminator to generate new training samples with diverse and plausible poses.
- We propose Pose Clustering Module (PCM) to measure the pose rarity and select rare poses for data augmentation, which helps to balance the long-tailed distribution of the training set.
- Extensive experiments on various pose estimation datasets show that PoseTrans consistently improves the performance of various state-of-the-art pose estimators, especially on rare poses.

2 Related Works

2.1 2D Human Pose Estimation

In recent years, 2D human pose estimation has shown remarkable performance advancement. DeepPose [41] first applied deep neural networks to human pose estimation by directly regressing the 2D coordinates of key points from the input image. Since then, deep learning-based methods started to dominate this area. Recent multi-person human pose estimation approaches can be divided into bottom-up and top-down approaches. Bottom-up approaches [23,7,36,34,25,13,28,24] first detect all the key points of every person in images and then group them into individuals. Top-down methods [21,11,45,40] first detect the bounding boxes and then predict the human body key points in each box.

Recent works mainly focus on designing powerful network architectures to improve the performance of pose estimation [35,45,40,11,26,46,48]. However, current state-of-the-art models often suffer performance drops on rare poses due to the long-tailed distribution problem in human pose data. In this work, we focus on tackling this important but ignored problem. Standing on the shoulder of the well-designed network structure, we propose a novel data augmentation method to generate diverse rare poses.

2.2 Data Augmentation

Data augmentation has been widely utilized to improve the model generalization ability. For image classification, popular augmentation methods include informa-

tion dropping [52,9,16], multi-image information mixing [50,47] and automatic augmentation [15]. For human pose estimation, data augmentation mainly focus on global image-level transformations [37,12,35,45,42] (*e.g* scaling, rotating, and flipping) and local object-level transformations [5,37] (*e.g* copy-paste, occluding). These common data augmentation schemes enhance the global translational invariance and robustness in occlusion cases but struggle to improve the immunity to rare poses. Recently, some augmentation methods [18,19] propose to perform jitting on instances to increase the generalization of the model, but they do not change either the instance itself or the distribution of instances. Different from the existing data augmentation strategies, we propose a novel, simple and effective PoseTrans augmentation scheme that directly generates diverse rare poses.

2.3 Long-tailed Distribution

In visual recognition, there exists a challenging problem of long-tailed training set distributions, where a small portion of classes have massive training samples while classes in the distribution tail have few samples [51]. Over-sampling [8] and re-weighting [17] are two popular methods to tackle the problem. The over-sampling method raises the frequency level of the minor classes by repeating the data samples during training. The re-weighting method assigns higher loss weights to these minor classes and thus increases their importance. However, such approaches do not increase the diversity of the data and tend to suffer from over-fitting which leads to a performance drop. Other approaches also include metric learning that enforces inter-class margins [22] and meta-learning that learns to regress many-shot model parameters from few-shot model parameters [44], but they are only designed for visual recognition. In human pose estimation, we encounter a similar problem. For many human pose estimation datasets [31,3,29], *e.g* the MS-COCO dataset [31], the distribution of human poses is highly biased, which does not uniformly represent human poses in real life. These dataset biases lead to poor generalization and degraded detection accuracy of these "long-tailed" poses. To address the aforementioned issue, we propose a simple yet effective PoseTrans approach to create the needed diverse poses.

3 Method

3.1 Overview

To increase the diversity of poses and alleviate the long-tailed distribution problem, we propose the Pose Transformation (PoseTrans) to generate new training samples with diverse poses, as shown in Fig. 2. PoseTrans consists of a Pose Transformation Module (PTM) with a pose discriminator D and a Pose Clustering Module (PCM). Given a training sample $(\boldsymbol{x}, \boldsymbol{y})$ consisting of a single human image $\boldsymbol{x}$ and its keypoint annotation $\boldsymbol{y}$, PTM aims to create a new training sample $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ by applying affine transformations on the limbs of the human, where $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{R}^{H\times W\times 3}$, $\boldsymbol{y}, \tilde{\boldsymbol{y}} \in \mathbb{R}^{J\times 2}$. H, W and J indicate the height, width and the

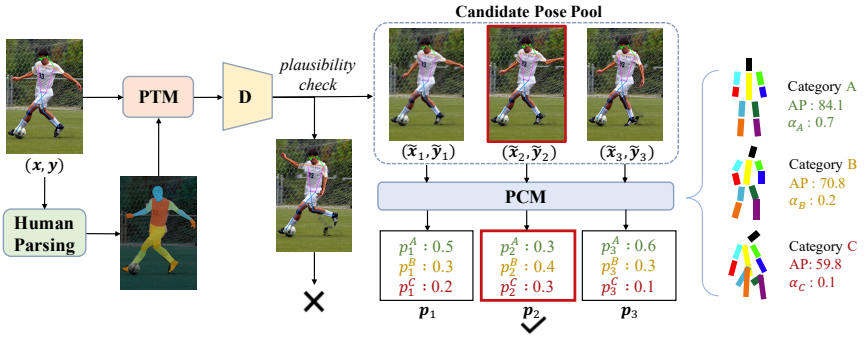

Fig. 2: Overview of PoseTrans. Given a single human image $\boldsymbol{x}$ and its keypoint annotations $\boldsymbol{y}$, we first segment the human into different parts through human parsing. PTM applies affine transformations on the limbs of the human to construct new poses. A pre-trained pose discriminator is used for the plausibility check. The plausible poses form a candidate pose pool $\{(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t)\}$, where $t \in \{1, 2, 3\}$ as an example. For pose $\tilde{\boldsymbol{y}}_t$, PCM predicts $\boldsymbol{w}_t$, which is the probability of belonging to each category (3 categories as an example). PCM selects the rarest one with the minimal weighted sum of components' density as a new training sample, *i.e* $w_2^A\alpha_A + w_2^B\alpha_B + w_2^C\alpha_C$.

number of keypoints respectively. To ensure plausibility, we leverage the pose discriminator D to filter out implausible samples. PoseTrans applies PTM repeatedly until a candidate pose pool with T plausible generated poses is formed. PCM clusters human poses into N categories and evaluates the probability of belonging to each cluster for generated poses to select the rarest one among the pool as a new training sample. After each training epoch, we re-fit the PCM using the original training set and all the selected augmented samples.

3.2 Pose Transformation Module (PTM) and Pose Discriminator

By clustering the human poses in the existing dataset, it can be observed that many clusters only have a few examples. The lack of training examples of rare poses further leads to the lack of diversity of rare poses, which results in the inferior performance of current data-driven methods on these types of poses. To tackle this issue, we devise the Pose Transformation Module (PTM) and a pose discriminator to create plausible new poses based on the existing training samples. The detail of PTM is shown in Fig. 3.

Modeling the body part movement. The body kinematic skeleton is constructed by a pose graph, where the human body is partitioned into several parts, *i.e* the head, the torso, the left/right arm, and the left/right leg. In this work, we mainly focus on the angular movement of the arms and legs. Angular movements (flexion and extension) take place at the shoulder, hip, elbow, knee, and wrist. Flexion decreases the angle between the bones (bending

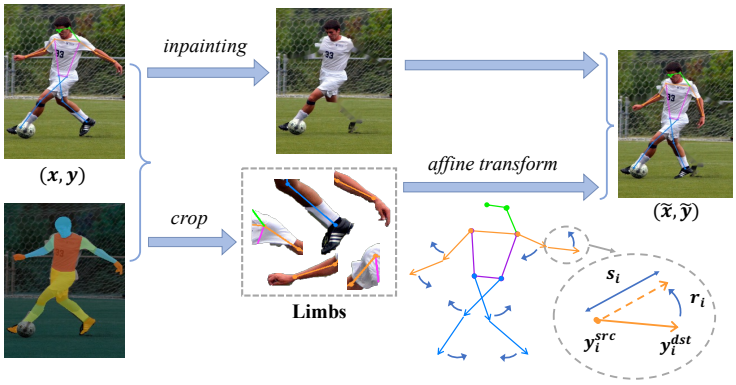

Fig. 3: By leveraging the human parsing results, we first erase the limbs from $\boldsymbol{x}$ and then transform each limb separately with a given probability $p = 0.5$. Limbs that do not appear or are obscured will not be transformed. The zoom-in view in the bottom right corner indicates the affine transformation with scale s_i and rotation r_i applied on the i-th limb (lower arm).

of the joint), while extension increases the angle and straightens the joint. These body part movements in the image plane can be modeled by applying the affine transformation to a rigid body part segment. In our implementation, the affine transformation is composed of rotation and scaling.

We define the limb as a single rigid body part connecting natural adjacent joints $\boldsymbol{y}^{src}$ and $\boldsymbol{y}^{dst}$, where $\boldsymbol{y}^{src}, \boldsymbol{y}^{dst} \in \mathbb{R}^2$ are the coordinates of the source and destination joint respectively. We define $K = 8$ limbs for each instance, including the lower arm, the upper arm, the lower leg, and the upper leg of both sides.

Pose transformation. With human parsing results obtained through DensePose [1] model, PTM first erases the original limbs in $\boldsymbol{x}$ by an efficient inpainting method [4]. After that, each limb is transformed by its affine transformation matrix separately. To increase the diversity, each limb has a probability of $p = 0.5$ to decide whether to transform or not. The transformed limbs and the inpainted image are composed to form the new augmented image $\tilde{\boldsymbol{x}}$. And the pose annotations are also transformed accordingly to get $\tilde{\boldsymbol{y}}$.

Specifically, the angular movement of the i-th limb can be modeled by the following affine transformation matrix

$$\boldsymbol{H}_i = \begin{bmatrix} s_i \cos r_i & -s_i \sin r_i & (1 - \cos r_i)c_i^x + c_i^y \sin r_i \\ s_i \sin r_i & s_i \cos r_i & (1 - \cos r_i)c_i^y - c_i^x \sin r_i \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where $s_i \in \mathbb{R}^+$ and $r_i \in \mathbb{R}$ denote the scale and rotation of the i-th limb, $\boldsymbol{y}_i^{src} = \{c_i^x, c_i^y\}$ is the coordinates of the rotation center of the i-th limb. For the lower arm, the upper arm, the lower leg, and the upper leg, the rotation centers are the elbow, the shoulder, the knee, and the hip respectively. To ensure the diversity of augmented poses, the scale s_i and rotation r_i parameters in $\boldsymbol{H}_i$ are

randomly sampled from a normal distribution in the neighboring space of identity transformation $(1, 0)$. The scale and rotation parameters are also restricted to a certain range in our implementation to ensure that the majority of the randomly generated poses are plausible. Note that, limbs that do not appear in the image or are obscured will not be transformed.

According to the kinematic skeleton hierarchy, the movement of the upper arm/leg will affect that of its lower part. Suppose the j-th limb is the lower arm/leg and the k-th limb is its corresponding upper part. Considering the combined effect, the total movement of the j-th limb can be modeled by matrix multiplication, *i.e* $\boldsymbol{H}_k\boldsymbol{H}_j$.

Pose discriminator for the plausibility check. Purely generating poses randomly may result in implausible poses that violate the biomechanical structure of the human body. Some other augmentation methods [30,10] rely on predefined rules for ensuring plausibility, which however limits the diversity of generated poses. Inspired by [20], we design a pose discriminator D that suits our task to avoid implausible poses that have unnatural joint angles or unreasonable positions in the scene. For the augmented sample $(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t)$, the discriminator D is trained to predict the plausibility $e_t = D\left(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t\right)$. We adopt the LS-GAN loss [33] to train the discriminator before training the pose estimatior:

$$\mathcal{L}_D = \mathbb{E}\left[\left(D(\boldsymbol{x}, \boldsymbol{y}) - 1\right)^2\right] + \mathbb{E}\left[D\left(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}\right)^2\right]. \tag{2}$$

With the pre-trained discriminator D, PoseTrans efficiently filter out the augmented sample whose plausibility is less than a pre-defined threshold $E \in [0, 1]$, and fill the candidate pose pool with samples that are plausible and diverse.

3.3 Pose Clustering Module (PCM)

After gaining the ability to create new human poses by PTM, we propose the Pose Clustering Module (PCM) to measure the pose rarity and select the needed poses for data augmentation.

Fitting the PCM. Our PCM is built upon the Gaussian Mixture Model (GMM) with N Gaussian components. As a soft clustering method, it predicts the probability of belonging to a certain category. Before pose clustering, human poses in the training set are first normalized. We crop every human instance on the image and re-scale the cropped image into the same height and width (256×256). The corresponding keypoint coordinates are also normalized at the same time. We fit the PCM using the normalized human poses in the training set. After fitting, given the pose $\boldsymbol{y}$, we model $P(\boldsymbol{y})$ as:

$$P(\boldsymbol{y}) = \sum_{n=1}^{N} \alpha_n \mathcal{N}\left(\boldsymbol{y}; \mu_n, \sigma_n\right), \tag{3}$$

where α_n is the weight of the n-th Gaussian component, $\mathcal{N}\left(\boldsymbol{y}; \mu_n, \sigma_n\right)$ denotes the n-th Gaussian distribution with mean μ_n and covariance σ_n.

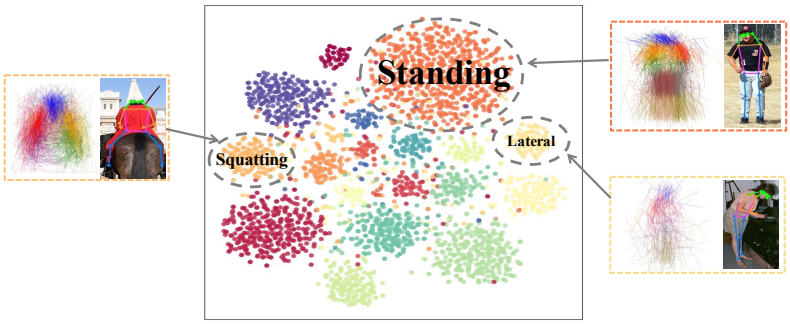

Fig. 4: The visualization of the clustering results using PCM by t-SNE. Different colored points indicate different clusters. Representative images and mean skeletons for the clusters of standing, squatting, and lateral poses are also visualized.

By predicting the probability of belonging to each Gaussian component, the human pose is classified as the component with the maximum probability. We visualize the probability vectors of every example using t-SNE [32], as shown in Fig. 4. With PCM, we cluster the human poses into N categories, where Gaussian components that have small weights (*i.e* few examples,) indicate the categories of rare poses. We observe the long-tailed problem that frontal standing accounts for a significant portion while squatting and lateral postures account for small percentages.

Pose selection from the candidate pose pool. PoseTrans repeats PTM to build a candidate pose pool $\{(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t)\}$ with T samples for the training sample $(\boldsymbol{x}, \boldsymbol{y})$, where $t \in \{1, 2, ..., T\}$. PoseTrans select the rarest one $(\tilde{\boldsymbol{x}}_{t^*}, \tilde{\boldsymbol{y}}_{t^*})$ among the candidate pose pool by:

$$t^* = \underset{t}{\operatorname{argmin}} \left(\sum_{n=1}^{N} \alpha_n w_t^n \right), \tag{4}$$

where $\boldsymbol{w}_t = \{w_t^1, w_t^2, \ldots, w_t^N\}$ is the predicted probability of $\tilde{\boldsymbol{y}}_t$ belonging to N Gaussian components by the fitted PCM. We consider the transformed sample $(\tilde{\boldsymbol{x}}_{t^*}, \tilde{\boldsymbol{y}}_{t^*})$ with the minimal weighted sum of components' density as the rarest and select it as a new training sample.

4 Experiments

4.1 Datasets and Evaluation

Datasets. To verify the effectiveness of our proposed data augmentation approach, we conduct extensive experiments on popular datasets. **(1)** MS-COCO [31] pose estimation dataset. Our models are trained on the train set only and evaluated on the val set and the test-dev set. DensePose [1] provides a small

portion of human parsing annotations for the MS-COCO dataset. To verify the performance on rare poses, both the traditional evaluation metrics (*i.e* AP/AR) and newly designed metrics (balanced AP/AR) are used for evaluation. The base learning rate of 1e-3, and decay the learning rate to 1e-4 and 1e-5 at the 170-th and 200-th epochs respectively. **(2)** PoseTrack'18 [2] dataset. Following common settings [14], we pre-train the model on the MS-COCO dataset and fine-tune it on the PoseTrack'18 dataset for 20 epochs. The basic learning rate is 1e-4 and drops to 1e-5 at 10 epochs then 1e-6 at 15 epochs. We test the model on the PoseTrack'18 validation set using the ground truth bounding boxes, and evaluate the AP on the whole body and also on different parts of the human. Due to the limited space, the results of some experiments are placed in the supplementary material.

Evaluation metrics. We follow [31] to use Average Precision (AP) and Average Recall (AR) for evaluation on MS-COCO [31]. They are based on object keypoint similarity (OKS), which measures the distance between predicted keypoints and ground-truth keypoints normalized by the scale of the object. AP_{50} (AP at OKS = 0.5), AP_{75} (AP at OKS = 0.75), AP^M for medium objects, and AP^L for large objects are reported.

Balanced AP/AR. Since existing datasets mostly suffer the long-tailed distribution problem, simply calculating the AP/AR tends to ignore the minor pose categories. To solve this problem, we design the balanced AP/AR, which we term AP_{BAL}, AR_{BAL}. We first classify the ground-truth poses into categories based on the fitted PCM. Then we calculate the standard AP/AR separately for each category and calculate the average precision/recall among *categories* instead of *samples*. Therefore, AP_{BAL} and AR_{BAL} assign the same weights to all pose categories, which is helpful to analyze the "unbiased" performance.

4.2 Implementation Details

PoseTrans can be integrated into the training pipeline of any existing pose estimators together with other common data augmentation strategies. Except for the small portion of images that have human parsing annotations, we leverage DensePose [1] model for human parsing which segments humans into 14 semantic parts. In PCM, we have $N = 20$ and cluster the poses into 20 categories. We implement PoseTrans with scaling ($s \in [0.75, 1.25]$), rotating ($r \in [-35°, 35°]$), and apply it with the probability $p = 0.5$ for every limb in the training examples. We filter out the implausible samples whose plausibility is less than $E = 0.7$ and repeat PTM until the candidate pose pool has $T = 5$ augmented samples.

For *bottom-up* methods, PoseTrans is applied on every instance in the image separately. The experimental settings are the same as [13]. We apply image-level random scaling ($[-25\%, 25\%]$), random rotation ($[-30°, 30°]$), random translation ($[-40\text{px}, 40\text{px}]$) and random flipping. The models are trained for 300 epochs using the Adam optimizer [27]. The base learning rate is 1e-3, and it decreases to 1e-4 and 1e-5 at the 200-th and 260-th epochs respectively. For *top-down* approaches, the experimental settings are the same as [40]. We use the detected

Table 1: **Improvements** on MS-COCO `val` set and `test-dev` set. PoseTrans consistently boosts the performance of the state of the arts.

Method	Input size	MS-COCO `val`						MS-COCO `test-dev`					
		AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Bottom-up methods w/o multi-scale test													
AE[34] + HRNet-W32[40]	512 × 512	64.4	86.3	72.0	57.1	75.6	71.0	64.1	86.3	70.4	57.4	73.9	70.4
+ PoseTrans (Ours)	512 × 512	**66.2**	**86.4**	**72.1**	**59.3**	**76.5**	**71.6**	**65.4**	**87.6**	**72.1**	**58.8**	**74.7**	**71.0**
HigherHRNet-W32[13]	512 × 512	67.1	86.2	73.0	61.5	76.1	72.3	66.4	87.5	72.8	61.2	74.2	71.4
+ PoseTrans (Ours)	512 × 512	**68.4**	**87.1**	**74.8**	**62.7**	**77.1**	**72.9**	**67.4**	**88.3**	**73.9**	**62.1**	**75.1**	**72.2**
Bottom-up methods with multi-scale test [×2, ×1, ×0.5]													
AE[34] + HRNet-W32[40]	512 × 512	68.5	87.1	75.1	64.0	76.8	73.9	68.1	88.3	75.1	63.8	74.9	72.9
+ PoseTrans (Ours)	512 × 512	**70.5**	**87.8**	**76.7**	**65.1**	**78.1**	**75.2**	**69.4**	**88.8**	**76.3**	**64.4**	**76.2**	**74.2**
HigherHRNet-W32[13]	512 × 512	69.9	87.1	76.0	65.3	77.0	74.7	68.8	88.8	75.7	64.4	75.0	73.5
+ PoseTrans (Ours)	512 × 512	**71.2**	**88.2**	**77.2**	**66.5**	**78.0**	**75.3**	**69.9**	**89.3**	**77.0**	**65.2**	**76.2**	**74.3**
Top-down methods													
SBL-ResNet-50[45]	256 × 192	70.4	88.6	78.3	67.1	75.9	76.3	70.2	90.9	78.3	67.1	75.9	75.8
+ PoseTrans (Ours)	256 × 192	**72.3**	**89.9**	**80.0**	**68.3**	**79.2**	**77.8**	**71.5**	**91.8**	**80.0**	**68.1**	**77.3**	**77.0**
SBL-ResNet-101[45]	256 × 192	71.4	89.3	79.3	68.1	78.1	77.1	71.1	91.5	79.6	67.7	76.8	76.6
+ PoseTrans (Ours)	256 × 192	**72.7**	**90.0**	**80.7**	**69.5**	**78.8**	**78.3**	**71.8**	**91.6**	**80.3**	**68.3**	**77.5**	**77.3**
HRNet-W32[40]	256 × 192	74.4	90.5	81.9	70.8	81.0	79.8	73.5	92.2	82.0	70.4	79.0	79.0
+ PoseTrans (Ours)	256 × 192	**75.5**	**91.0**	**82.9**	**71.8**	**82.2**	**80.7**	**74.2**	**92.4**	**82.5**	**70.8**	**79.6**	**79.4**
HRNet-W32[40] + Dark[49]	256 × 192	75.6	90.5	82.1	71.8	82.8	80.8	74.6	92.4	82.9	71.2	80.3	79.9
+ PoseTrans (Ours)	256 × 192	**76.0**	**90.8**	**83.0**	**72.1**	**83.2**	**81.1**	**75.0**	**92.5**	**82.9**	**71.5**	**80.6**	**80.1**
HRNet-W32[40]	384 × 288	75.8	90.6	82.7	71.9	82.8	80.1	74.9	92.5	82.8	71.3	80.9	80.1
+ PoseTrans (Ours)	384 × 288	**76.5**	**90.9**	**83.3**	**72.5**	**83.3**	**81.5**	**75.4**	**92.5**	**83.0**	**71.6**	**81.1**	**80.4**
HRNet-W48[40]	384 × 288	76.3	90.8	82.9	72.3	83.4	81.2	75.5	92.5	83.3	71.9	81.5	80.5
+ PoseTrans (Ours)	384 × 288	**76.8**	**91.0**	**83.1**	**72.7**	**83.7**	**81.6**	**75.7**	**92.6**	**83.4**	**72.0**	**81.7**	**80.6**

bounding boxes provided by Xiao *et al* [45]. The detection boxes are first extended to a fixed aspect ratio (*i.e* height:width = 4:3) and then enlarged by a factor of 1.25 to include some context. We apply random scaling ($[-35\%, 35\%]$), random rotation ($[-45°, 45°]$), random flipping, and half-body crops. The models are trained on 16 GPUs for 210 epochs. We use Adam optimizer [27] for training. All networks are pre-trained on the ImageNet dataset [39].

4.3 Improvement of state-of-the-art methods by PoseTrans

Improvement of AP/AR. Table 1 reports the performance improvement of AP and AR on the MS-COCO `val` and MS-COCO `test-dev` set, where PoseTrans is applied to recent state-of-the-art pose estimators, *i.e* SBL [45], HRNet [40], and HigherHRNet [13]. Table 2 show the performance improvement on the PoseTrack dataset. PoseTrans consistently boosts the performance of both top-down and bottom-up approaches in various datasets.

Improvement of $\mathbf{AP_{BAL}}$ and $\mathbf{AR_{BAL}}$. The results of AP_{BAL} and AR_{BAL} are reported in Table 3a. To calculate the new metrics, we use the bounding boxes and keypoint annotations to determine the category of predicted poses. Thanks to the design of PCM and PTM, PoseTrans increases the diversity of rare poses and balances the distribution, which enables PoseTrans to bring more improvements on the newly proposed AP_{BAL}/AR_{BAL} than traditional AP/AR.

Table 2: **Improvements** on PoseTrack2018 validation set.

Method	Input size	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total AP
SBL-ResNet-50 [45]	256 × 192	86.5	87.5	82.3	75.6	79.9	78.6	74.0	81.0
+ PoseTrans (Ours)	256 × 192	**87.8**	**89.3**	**84.7**	**77.7**	**82.3**	**81.6**	**75.4**	**83.0**
HRNet-W32 [40]	256 × 192	87.4	88.6	84.3	78.5	79.7	81.8	78.8	83.0
+ PoseTrans (Ours)	256 × 192	**88.6**	**90.0**	**86.2**	**80.3**	**83.1**	**84.9**	**79.8**	**84.9**
HRNet-W32 [40]	384 × 288	88.5	89.5	86.0	80.4	81.6	83.4	78.9	84.3
+ PoseTrans (Ours)	384 × 288	**88.9**	**90.3**	**87.4**	**81.8**	**83.5**	**85.5**	**80.6**	**85.7**

Table 3: **(a)** Improvements of Balanced AP/AR on MS-COCO `val` set. **(b)** Comparisons of data augmentation techniques on MS-COCO `val` set. HRNet-W32 with an input size of 256 × 192 is adopted as the baseline. Results marked with '*' are reported by [38] using CascadeRCNN bounding boxes.

Method	Input size	MS-COCO `val`			
		AP	AR	AP_{BAL}	AR_{BAL}
SBL-ResNet50 [45]	256 × 192	70.4	76.3	60.6	66.3
+ PoseTrans (Ours)	256 × 192	**72.3**	**77.8**	**63.8**	**69.6**
HRNet-W32 [40]	256 × 192	74.4	79.8	65.4	72.3
+ PoseTrans (Ours)	256 × 192	**75.5**	**80.7**	**67.9**	**73.8**
HRNet-W32 [49]	384 × 288	75.8	80.1	67.7	73.8
+ PoseTrans (Ours)	384 × 288	**76.5**	**81.5**	**68.9**	**74.2**

(a)

Method	AP	AP^{50}	AP^{75}	AR
Baseline [40]	74.4	90.5	81.9	79.8
+ Cutout* [16]	74.5	90.5	81.7	78.8
+ GridMask [9]	74.7	90.6	82.0	80.1
+ Photometric Distortion [6]	74.6	90.3	81.9	80.0
+ AdvMix [42]	74.7	-	-	-
+ InstaBoost [18]	74.7	90.5	82.0	80.1
+ ASDA [5]	75.2	**91.0**	82.4	80.4
+ PoseTrans (Ours)	**75.5**	**91.0**	**82.9**	**80.7**

(b)

4.4 Comparisons with other data augmentation techniques

In Table 3b, we compare PoseTrans with other data augmentation techniques, including non-learning [16,9,6] and learning/strategy-based methods [42,18].

For non-learning methods, Cutout [16] randomly selects a rectangle region around the keypoint and fills in random values. GridMask [9] evenly replaces multiple rectangle regions in an image with all zeros. For Photometric Distortion, we follow [6] to adjust the brightness, contrast, hue, saturation, and noise of an image. These general data techniques are proven to be effective for image classification. However, they do not bring significant improvements for human pose estimation. Similar conclusions have also been reached by previous works [38]. This is probably because such techniques introduce undesirable artifacts and do not increase the diversity of human poses.

For learning/strategy-based methods, AdvMix [42] applies adversarial training to learn to mix up augmented samples generated by GridMask [9] and AutoAugment [15]. InstaBoost [18] is a recently proposed data augmentation technique which is originally designed for instance segmentation. It conducts crop-paste augmentation guided by the appearance consistency heatmaps. However, the improvements of AdvMix and InstaBoost are only marginal. ASDA [5] also employs human parsing and augments images by pasting the segmented

body parts. PoseTrans outperforms all these approaches, which validates the importance of increasing the diversity of the human body poses.

Kindly note that PoseTrans is also complementary to other techniques. Effectively combining these techniques may further improve the final performance. As shown in the third row from the bottom in Table 1, combining PoseTrans with DarkPose [49] can further gain improvements.

4.5 Ablation Studies

Effect of PTM. Without using the PTM, we perform the over-sampling [8] and re-weighting [17] strategies, which are two popular methods to tackle the long-tailed problem. The over-sampling method raises the frequency level of the minor categories by duplicating the long-tailed data samples during model training. The re-weighting method assigns higher loss weights to rare samples and thus increases their importance. Based on the clustering results of PCM, we implement these methods as baselines, as shown in Table 4a. By increasing the importance of long-tailed training samples, both the over-sampling marginally improve the $\mathrm{AP}_{\mathrm{BAL}}$. However, such approaches do not increase the diversity of the data, which leads to slight performance drops on AP and AR. With the design of PTM, our proposed PoseTrans creates diverse long-tailed samples, which significantly outperforms the baseline methods.

Effect of PCM. Without PCM, PoseTrans randomly samples a transformed pose obtained from PTM as the training sample, instead of picking the "rarest" pose in the candidate pose pool. Note that, "w/o PCM" is equivalent to the case of $T = 1$ in PoseTrans. The studies of w/o PCM and the number of T in PCM are shown in Table 4b. By providing simple disturbance to training data, w/o PCM increases the generalization of the model, which leads to some performance improvements. While with the aid of PCM, our full model learns to alleviate the long-tailed distribution problem of the training set by selecting transformed poses, which brings greater performance gains, especially on $\mathrm{AP}_{\mathrm{BAL}}/\mathrm{AR}_{\mathrm{BAL}}$. Also, a larger candidate pose pool (*i.e* greater T) leads to better performance. However, T greater than 5 will not bring more performance boost.

Effect of pose discriminator. Without the pose discriminator (D), some implausible poses will lead to performance degradation as shown in Table 5a. Since the scale s and rotation r parameters are sampled from a normal distribution and are restricted to $[0.75, 1.25]$ and $[-35°, 35°]$ in the implementation, a majority of the randomly generated poses are plausible. In this situation, the PTM without the pose discriminator can still benefit the model.

Comparison with the adversarial learning variant. Inspired by recent works [42,37,5] on adversarial data augmentation, we also build an adversarial training variant of PoseTrans, which we refer to as PoseTrans-Adv. PoseTrans-Adv has an additional generator that predicts the rotation r and scale s for a given single human image $\boldsymbol{x}$. During training, the generator is asked to confuse the pose estimation model by maximizing the loss of the pose estimator. However, we observe that the generator will soon learn to choose the maximum rotation and scale for every training sample, which actually decreases the diversity of the

Table 4: **(a)** Ablation studies of PTM. The over-sampling and re-weighting methods are based on the clustering results of PCM. **(b)** Ablation studies of PCM. HRNet-W32 with the input size of 256×192 is adopted for experiments.

Method	AP	AR	AP_{BAL}	AR_{BAL}
Baseline	74.4	79.8	65.4	72.3
Over-sampling [8]	74.3	79.7	66.0	72.3
Re-weighting [17]	74.2	79.6	65.8	72.2
PoseTrans (Ours)	**75.5**	**80.7**	**67.9**	**73.8**

(a)

Method	AP	AR	AP_{BAL}	AR_{BAL}
Baseline [40]	74.4	79.8	65.4	72.3
w/o PCM	74.9	80.1	66.1	72.6
PoseTrans ($T = 3$)	75.2	80.3	67.2	72.9
PoseTrans ($T = 5$)	**75.5**	**80.7**	**67.9**	**73.8**

(b)

Table 5: **(a)** Ablation studies of Discriminator (D). **(b)** Comparison with the variants of PoseTrans.

Method	AP	AR	AP_{BAL}	AR_{BAL}
PoseTrans w/o D	75.0	80.1	66.5	72.8
PoseTrans	**75.5**	**80.7**	**67.9**	**73.8**

(a)

Method	AP	AR	AP_{BAL}	AR_{BAL}
PoseTrans-Adv	72.7	78.4	65.2	71.5
PoseTrans-Par	75.3	80.4	67.3	73.3
PoseTrans	**75.5**	**80.7**	**67.9**	**73.8**

(b)

training set. This leads to performance degradation in all the evaluation metrics as shown in the first row of Table 5b.

Comparison with PoseTrans-Par on the MS-COCO dataset. As mentioned above, DensePose [1] provides a small portion of human parsing annotations for the MS-COCO dataset. Here, we compare with the PoseTrans-Par variant that replaces the human annotations with the pseudo-labels obtained from the parsing model. As shown in the second row of Table 5b, without human annotations, the performance of PoseTrans-Par is comparable with PoseTrans.

4.6 Analysis

Visualizations of the augmented samples. In Fig. 5, we visualize the original image and the augmented sample by PoseTrans. It can be observed that our proposed method generates diverse and plausible body postures that facilitate the model training and improve its generalization ability.

Visualizations of pose estimation results. In Fig. 6, we visualize pose estimation results obtained by HRNet [40]. We observe that vanilla HRNet is easily confused by infrequent and difficult poses, *e.g* upside-down postures and serious occlusions. By generating training samples with diverse rare poses, our PoseTrans improves the performance in these challenging cases.

Limitations. Our limitations mainly lie in the artifacts produced by the inpainting method and the accuracy of the human parsing model. We choose a simple non-data-driven inpainting method in pose transformation for efficiency. An advanced inpainting and parsing model with higher resolution inputs may bring more improvements in pose estimation.

Fig. 5: **Visualizations** of PoseTrans augmented samples. We observe that our proposed method generates more diverse body postures which facilitates the model training and improves its generalization ability.

Fig. 6: **Qualitative comparisons** of vanilla HRNet [40] (upper row) and HRNet trained with PoseTrans (bottom row). PoseTrans improves the human pose estimation results, especially for rare poses.

5 Conclusions

In this paper, we study the performance degradation caused by unbalanced data distribution on human pose estimation. To tackle this issue, we propose PoseTrans with PTM, PCM, and a pose discriminator to create diverse and plausible training samples that have infrequent poses. Comprehensive experiments on public benchmarks demonstrate the effectiveness of our method, especially on rare poses. Our implementation of PoseTrans is simple and efficient, which can be easily integrated into the training pipeline of existing pose estimators. We hope our work will draw the community's attention to the long-tail problem in human pose estimation and provide inspiration on how to tackle it for other tasks.

Acknowledgement. This work is supported in part by the National Natural Science Foundation of China under Grant 62122010 and Grant 61876177, in part by the Fundamental Research Funds for the Central Universities, and in part by the Key Research and Development Program of Zhejiang Province under Grant 2022C01082. Ping Luo is supported by the General Research Fund of HK No.27208720, No.17212120, and No.17200622.

References

1. Alp Güler, R., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018) 6, 8, 9, 13
2. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5167–5176 (2018) 9
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2014) 4
4. Bertalmio, M., Bertozzi, A.L., Sapiro, G.: Navier-stokes, fluid dynamics, and image and video inpainting. In: IEEE Conf. Comput. Vis. Pattern Recog. vol. 1, pp. I–I. IEEE (2001) 6
5. Bin, Y., Cao, X., Chen, X., Ge, Y., Tai, Y., Wang, C., Li, J., Huang, F., Gao, C., Sang, N.: Adversarial semantic data augmentation for human pose estimation. In: Eur. Conf. Comput. Vis. pp. 606–622. Springer (2020) 2, 4, 11, 12
6. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020) 11
7. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017) 3
8. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002) 4, 12, 13
9. Chen, P., Liu, S., Zhao, H., Jia, J.: Gridmask data augmentation. arXiv preprint arXiv:2001.04086 (2020) 4, 11
10. Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., Lischinski, D., Cohen-Or, D., Chen, B.: Synthesizing training images for boosting human 3d pose estimation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 479–488. IEEE (2016) 7
11. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018) 3
12. Chen, Y., Shen, C., Wei, X.S., Liu, L., Yang, J.: Adversarial posenet: A structure-aware convolutional network for human pose estimation. In: Int. Conf. Comput. Vis. pp. 1212–1221 (2017) 2, 4
13. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5386–5395 (2020) 3, 9, 10
14. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020) 9
15. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018) 4, 11
16. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017) 4, 11
17. Elkan, C.: The foundations of cost-sensitive learning. In: International joint conference on artificial intelligence. vol. 17, pp. 973–978. Lawrence Erlbaum Associates Ltd (2001) 4, 12, 13

18. Fang, H.S., Sun, J., Wang, R., Gou, M., Li, Y.L., Lu, C.: Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In: Int. Conf. Comput. Vis. pp. 682–691 (2019) 2, 4, 11
19. Fang, H.S., Xie, Y., Shao, D., Li, Y.L., Lu, C.: Decaug: Augmenting hoi detection via decomposition. In: AAAI. pp. 1300–1308 (2021) 4
20. Gong, K., Zhang, J., Feng, J.: Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8575–8584 (2021) 7
21. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. arXiv preprint arXiv:1703.06870 (2017) 3
22. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5375–5384 (2016) 4
23. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: Eur. Conf. Comput. Vis. (2016) 3
24. Jin, S., Liu, W., Ouyang, W., Qian, C.: Multi-person articulated tracking with spatial and temporal embeddings. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019) 3
25. Jin, S., Liu, W., Xie, E., Wang, W., Qian, C., Ouyang, W., Luo, P.: Differentiable hierarchical graph grouping for multi-person pose estimation. In: Eur. Conf. Comput. Vis. pp. 718–734 (2020) 3
26. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Eur. Conf. Comput. Vis. pp. 196–214 (2020) 3
27. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9, 10
28. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11977–11986 (2019) 3
29. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10863–10872 (2019) 4
30. Li, S., Ke, L., Pratama, K., Tai, Y.W., Tang, C.K., Cheng, K.T.: Cascaded deep monocular 3d human pose estimation with evolutionary training data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6173–6183 (2020) 7
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Eur. Conf. Comput. Vis. (2014) 2, 4, 8, 9
32. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 8
33. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2017) 7
34. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Adv. Neural Inform. Process. Syst. (2017) 1, 3, 10
35. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Eur. Conf. Comput. Vis. (2016) 1, 2, 3, 4

36. Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. arXiv preprint arXiv:1803.08225 (2018) 3
37. Peng, X., Tang, Z., Yang, F., Feris, R.S., Metaxas, D.: Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2226–2234 (2018) 2, 4, 12
38. Pytel, R., Kayhan, O.S., van Gemert, J.C.: Tilting at windmills: Data augmentation for deep pose estimation does not help with occlusions. In: Int. Conf. Pattern Recog. pp. 10568–10575 (2021) 11
39. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015) 10
40. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. arXiv preprint arXiv:1902.09212 (2019) 3, 9, 10, 11, 13, 14
41. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1653–1660 (2014) 1, 3
42. Wang, J., Jin, S., Liu, W., Liu, W., Qian, C., Luo, P.: When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11855–11864 (2021) 2, 4, 11, 12
43. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2020) 2
44. Wang, Y.X., Ramanan, D., Hebert, M.: Learning to model the tail. In: Adv. Neural Inform. Process. Syst. pp. 7032–7042 (2017) 4
45. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Eur. Conf. Comput. Vis. (2018) 2, 3, 4, 10, 11
46. Xu, L., Guan, Y., Jin, S., Liu, W., Qian, C., Luo, P., Ouyang, W., Wang, X.: Vipnas: Efficient video pose estimation via neural architecture search. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021) 3
47. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Int. Conf. Comput. Vis. pp. 6023–6032 (2019) 4
48. Zeng, W., Jin, S., Liu, W., Qian, C., Luo, P., Ouyang, W., Wang, X.: Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022) 3
49. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7093–7102 (2020) 10, 11, 12
50. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017) 4
51. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. arXiv preprint arXiv:2110.04596 (2021) 4
52. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI. vol. 34, pp. 13001–13008 (2020) 4