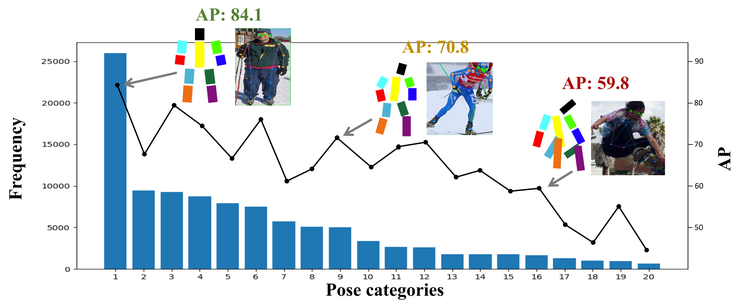
PoseTrans: A Simple Yet Effective Pose Transformation Augmentation for Human Pose Estimation

Wentao Jiang[1,2], Sheng Jin[2,3], Wentao Liu[2,4], Chen Gao[1], Chen Qian[2], Ping Luo[3], Si Liu[1]
[1]Beihang University, [2]SenseTime Research, [3]University of Hong Kong, [4]Shanghai AI Lab

ECCV TEL AVIV 2022

Motivation

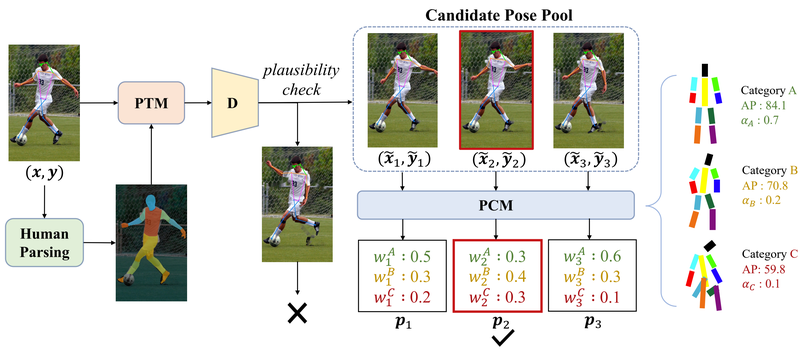
We cluster the poses in the MS-COCO dataset into 20 categories and evaluate the AP. The top-1 category has more than 25000 samples and high precision, while nearly half of the categories have less than 2000 samples and relatively low precision.



Pose estimators suffer performance degradation on some unusual poses, since the long-tailed categories have neither enough training samples nor enough diversity.

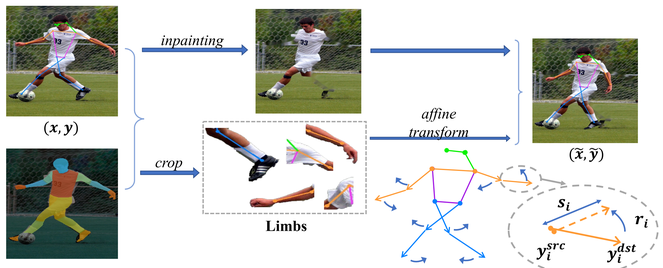
Framework

PoseTrans consists of a Pose Transformation Module (PTM) with a pose discriminator D and a Pose Clustering Module (PCM). Given a training sample, PTM aims to create a new training sample by applying affine transformations on the limbs of the human. To ensure plausibility, we leverage the discriminator D to filter out implausible samples.



PoseTrans applies PTM repeatedly until a candidate pose pool with T plausible 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.

Pose Transformation Module (PTM)



PTM

By leveraging the human parsing results, we first erase the limbs from image and then transform each limb separately with a given probability p. 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)

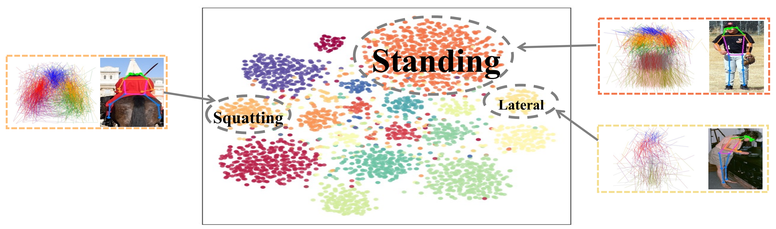
Pose Discriminator D

We design the pose discriminator D to avoid implausible poses that have unnatural joint angles or unreasonable positions in the scene. We adopt the LS-GAN loss to train the discriminator before training the pose estimator.

Pose Clustering Module (PCM)

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

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.



Visualization of the clusters using t-SNE. Different colored points indicate different clusters. Representative images and mean skeletons for the clusters of standing, squatting, and lateral poses are visualized.

Experimental Results

Comparisons

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` AP	AP^{50}	AP^{75}	AP^M	AP^L	AR	MS-COCO `test-dev` 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[46]	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[46]	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[48]	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**

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 [46]	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 [48]	384 × 288	75.8	80.1	67.7	73.8
+ PoseTrans (Ours)	384 × 288	**76.5**	**81.5**	**68.9**	**74.2**

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 [43]	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**

Ablation Studies

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**

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**

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**

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**