PoseAug: A Differentiable Pose Augmentation Framework for 3D Human Pose Estimation (Supplementary Material)

Kehong Gong [*] [†] Jianfeng Zhang [*] Jiashi Feng
National University of Singapore
gongkehong@u.nus.edu zhangjianfeng@u.nus.edu elefjia@nus.edu.sg

A. Implementation and training details

Implementation details We use the single-frame version of VPose [7] as our pose estimator. The pose augmentor consists of four linear layers with Batch Normalization (BN) [3] and leaky ReLU [2]. It first transforms the input 3D pose $\boldsymbol{X}$ to 256-D hidden features and then predicts the augmentation parameters $\boldsymbol{\gamma_{ba}}$, $\boldsymbol{\gamma_{bl}}$, $\boldsymbol{R}$ and $\boldsymbol{t}$ from the hidden features. Both the 3D and 2D pose discriminators consist of 4 linear layers with leaky ReLU and residual connection. The hidden feature dimensions are set as 256 and 100 for 3D and 2D discriminator, respectively. We train our model for 50 epochs on Human3.6M, with batch size of 1024. We adopt Adam optimizer [5] with linear decay and an initial learning rate of 0.001 for all components (*i.e.*, augmentor, discriminator, estimator). The hard ratio β linearly increases from 2 to 20 during the training process. The threshold for regularizing the augmentation parameters $\boldsymbol{\gamma_{ba}}$ and $\boldsymbol{\gamma_{bl}}$ is set as 0.1.

Training details With the differentiable augmentor design, the pose augmentor $\mathcal{A}$ with parameters $\boldsymbol{\theta}_A$, discriminator $\mathcal{D}$ with parameters $\boldsymbol{\theta}_D$ and estimator $\mathcal{P}$ with parameters $\boldsymbol{\theta}$ can be jointly trained end-to-end as illustrated in Algorithm 1. In detail, we firstly update the augmentor $\mathcal{A}$ and discriminator $\mathcal{D}$ alternatively by minimizing the pose augmentation loss Eqn. (11) and discrimination loss Eqn. (12) with the source pose pairs $\{\boldsymbol{x}_i, \boldsymbol{X}_i\}_{i=1}^N$, and augmented pose pairs $\{\boldsymbol{x'}_i, \boldsymbol{X'}_i\}_{i=1}^N$. We then use both the source and augmented pose pairs to train the estimator $\mathcal{P}$ by minimizing the pose estimation loss Eqn. (8).

Algorithm 1: PoseAug training strategy

Input: N training samples $\{\boldsymbol{x}_i, \boldsymbol{X}_i\}_{i=1}^N$, training epochs E and learning rate α.
Output: pose estimator $\mathcal{P}$, augmentor $\mathcal{A}$ and discriminator $\mathcal{D}$.
for $e = 1, \cdots, E$ **do**

// Update augmentor $\mathcal{A}$
Generate augmented sample $\{\boldsymbol{x'}_i, \boldsymbol{X'}_i\}_{i=1}^N$ from $\{\boldsymbol{x}_i, \boldsymbol{X}_i\}_{i=1}^N$ through the augmentor $\mathcal{A}$
Calculate the augmentation loss $\mathcal{L}_\mathcal{A}$ (Eqn. (11))
Calculate gradients $\nabla_{\boldsymbol{\theta}_A}\mathcal{L}_\mathcal{A}$ and update the parameters of $\mathcal{A}$ by $\boldsymbol{\theta}_A = \boldsymbol{\theta}_A - \alpha\nabla_{\boldsymbol{\theta}_A}\mathcal{L}_\mathcal{A}$

// Update discriminator $\mathcal{D}$
Calculate the discrimination loss $\mathcal{L}_\mathcal{D}$ (Eqn. (12))
Calculate gradients $\nabla_{\boldsymbol{\theta}_D}\mathcal{L}_\mathcal{D}$ and update the parameters of $\mathcal{D}$ by $\boldsymbol{\theta}_D = \boldsymbol{\theta}_D - \alpha\nabla_{\boldsymbol{\theta}_D}\mathcal{L}_\mathcal{D}$

// Update estimator $\mathcal{P}$
Calculate the estimation loss $\mathcal{L}_\mathcal{P}$ (Eqn. (8)) by feeding $\{\boldsymbol{x}_i, \boldsymbol{X}_i\}_{i=1}^N$ and $\{\boldsymbol{x'}_i, \boldsymbol{X'}_i\}_{i=1}^N$ alternatively to $\mathcal{P}$
Calculate gradients $\nabla_{\boldsymbol{\theta}}\mathcal{L}_\mathcal{P}$ and update the parameters of $\mathcal{P}$ by $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha\nabla_{\boldsymbol{\theta}}\mathcal{L}_\mathcal{P}$

end

B. More qualitative results

Here we present more qualitative results of the pose estimator (*i.e.*, VPose [7] (1-frame)) trained with and w/o our PoseAug on LSP [4], MPII [1], and MPI-INF-3DHP [6], as shown in Fig. S1. The comparison of inference results between with and w/o our PoseAug demonstrate that our auto-augmentation framework helps the pose estimator achieve better performance on challenge in-the-wild scenes. In addition, Fig. S2 shows three examples of source and corresponding augmented pose pairs. These examples demonstrate that our augmentor applies meaningful augmentation operation in generating harder cases, *e.g.*, harder posture (1st row), harder view point (2nd row), and even unseen pose (3rd row), which increases the data diversity and leads

[*]Equal contribution. [†]Work done during an internship at huawei Singapore research center.

to better generalization for estimator.

C. Ablation on feedback

Besides, the error feedback strategy brings significant improvement on 3DPW dataset (from 134.1 to 130.3, Table S1). This clearly verifies its effectiveness, especially on challenging scenarios.

Table S1: Ablation study on error feedback strategy on 3DPW. Augmentation denotes the combination of BA, BL, RT operations.

Method	Augmentation	Feedback	PA-MPJPE (↓)
Baseline			145.2
Variant A	✓		134.1 (-11.1)
PoseAug	✓	✓	**130.3** (-14.9)

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv*, 2015.

[4] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICCV*, 2015.

[6] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017.

[7] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019.

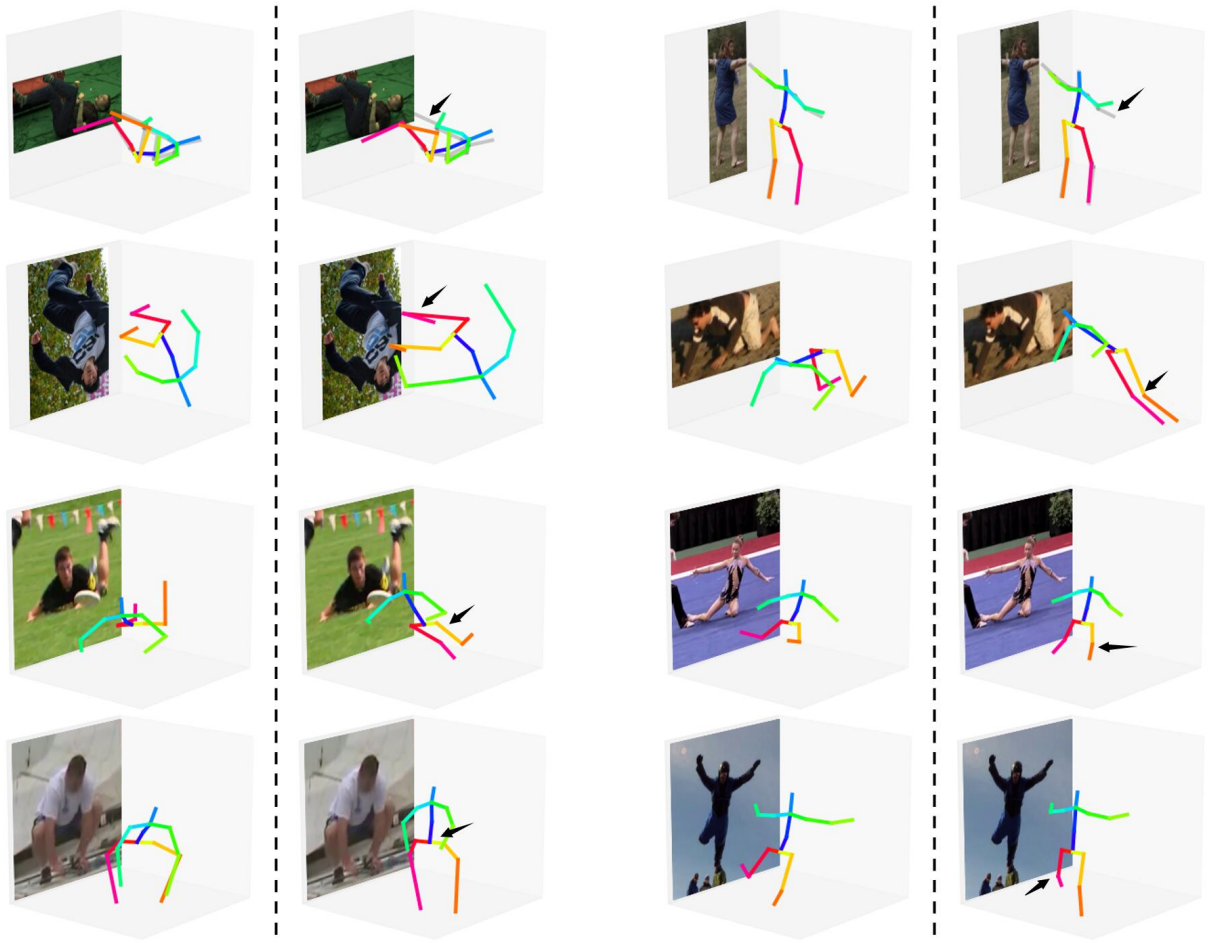

Figure S1: Example 3D pose estimations from 3DHP (1st row), LSP (2nd row), and MPII (3rd and 4th row). PoseAug results are shown in the 1st and 3rd columns. The 2nd and 4th columns show the results of *Baseline*, *i.e.*, VPose [7] (1-frame) trained without PoseAug. The grey skeletons in 3DHP (1st row) are ground truth. Errors are highlighted by black arrows.

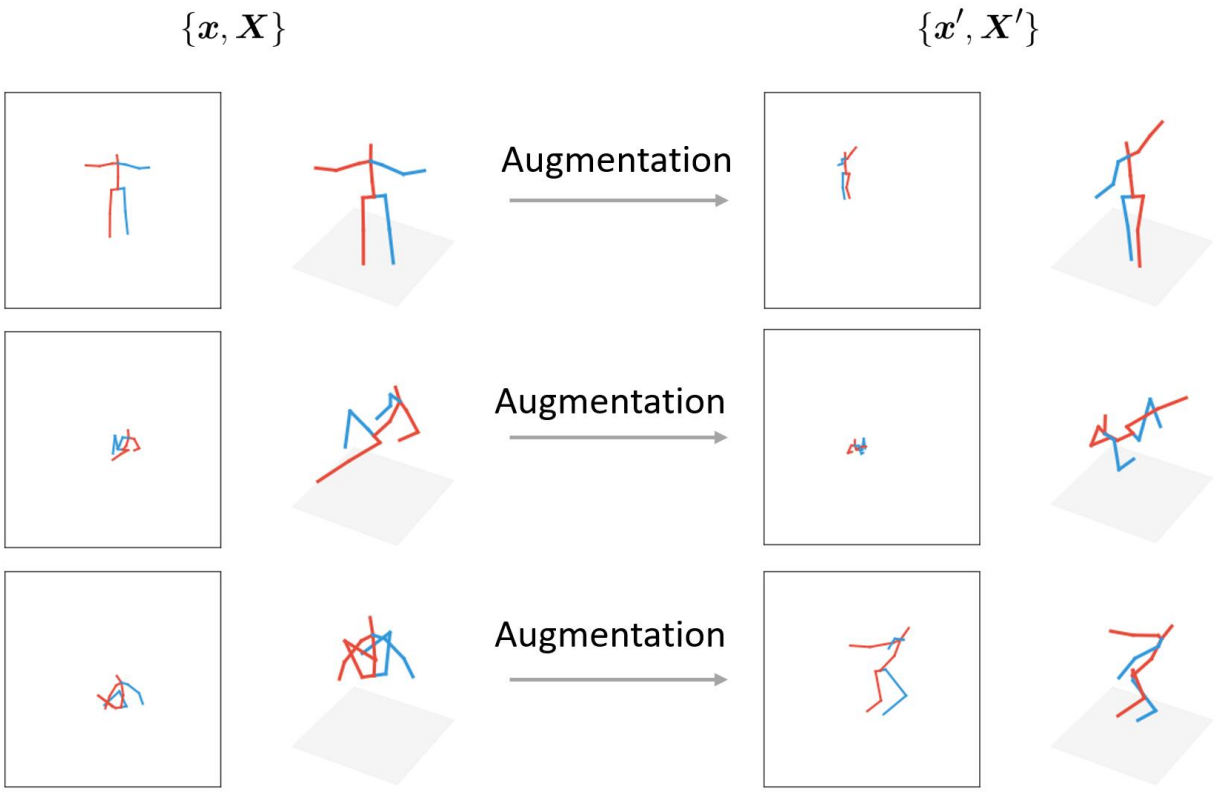

Figure S2: Examples of the source pose pair ($\boldsymbol{x}$: 1st column, $\boldsymbol{X}$: 2nd column) and its augmented pose pair ($\boldsymbol{x}'$: 3rd column, $\boldsymbol{X}'$: 4th column) . The examples include harder posture (1st row), harder view point (2nd row), and unseen pose (3rd row).