Supplementary Materials
A Synthesis Approach to Semi-Supervised 3D Pose Estimation from Single Images

Anonymous CVPR submission

Paper ID 1818

1. Encoder-Decoder Architecture

Our encoder is a standard ResNet34 [1], and the decoder is the inverse of ResNet by replacing all the downsampling by upsampling. Fully-connected layers are applied to reshape the tensor to the desired dimension. Specifically, the encoder consists of a convolution layer, five residual layers, and a fully-connected layer. The first convolution layer changes the image shape from $3 \times 256 \times 256$ to $64 \times 128 \times 128$. Then the five residual layers are applied and the tensor shape after each residual layer are respectively $64 \times 64 \times 64, 64 \times 32 \times 32, 128 \times 16 \times 16, 256 \times 8 \times 8, 512 \times 4 \times 4$. Average pooling is applied to the output of the last residual layer so that the data shape becomes 512. A fully connected layer is then applied to shrink the dimension to the embedding length (256 in this work).

The five residual layers are respectively the cascade of $3, 4, 6, 4, 3$ residual blocks (with equal input and output shape) with a downsampling convolution. The residual block has two 3×3 same-padding convolution layers. The output of the residual block is the sum of the input and output of the convolution layers given the input. Downsampling is a 1×1 convolution with stride 2. The structure of the decoder is obtained by replacing all the downsampling with upsampling: 4×4 convTranspose with stride 2. Figure 2 shows the architecture of the encoder and the decoder.

2. Synthesis of Training Data

During the experiments on H3.6M [2] and MPI-INF-3DHP [3], we enriched the 2D poses in the training set by convexification-sparsification. Specifically, for each joint (limb), we compute the relative position of the child node with respect to its parent node. Then we continuously randomly choose two relative positions and linearly combine them randomly. This step effectively convexifies the relative positions and the generated relative positions will be pooled into the training set. After the enlargement of the training set, we do random sampling on the enlarged training set so that the size is manageable. Figure 1 shows the relative position of the child nodes given the position of their parent node before and after the enrichment. After conducting the convexification-sparsification on every joint, we generate 2D poses by randomly sewing the joints and used [4] to generate the synthetic images.

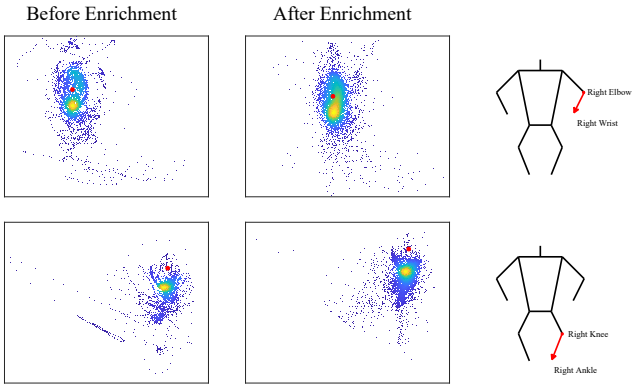

Figure 1. Relative positions of the child node given its parent node before and after our enrichment in the training set of MPI-INF-3DHP [3] (S1 Seq1 subset). The red dots are the parent nodes, which are respectively the right elbow and the right knee. The red arrow in the 2D model indicates the corresponding joint.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7), 2013. 1

[3] 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. 1

[4] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. 1

[5] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. doi:10.5244/C.24.12. 2

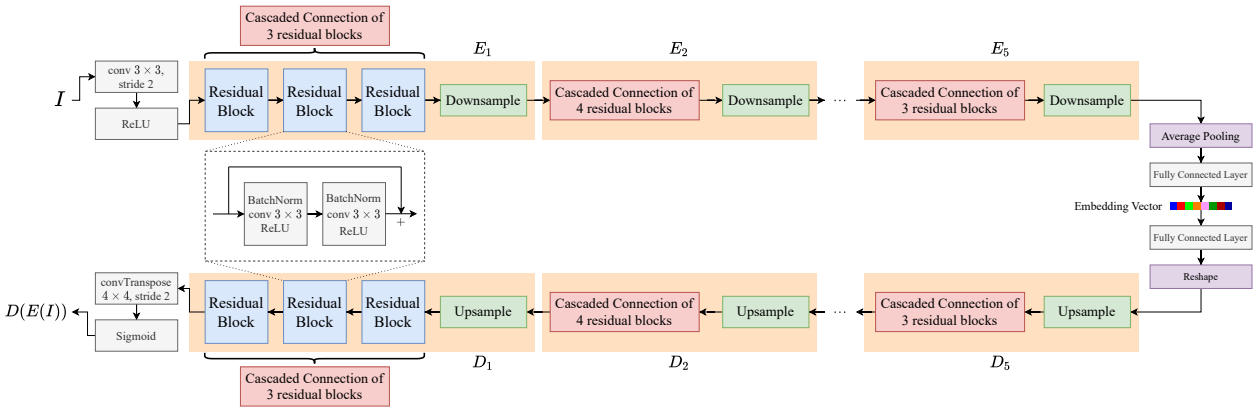

Figure 2. Architecture of the encoder and the decoder.

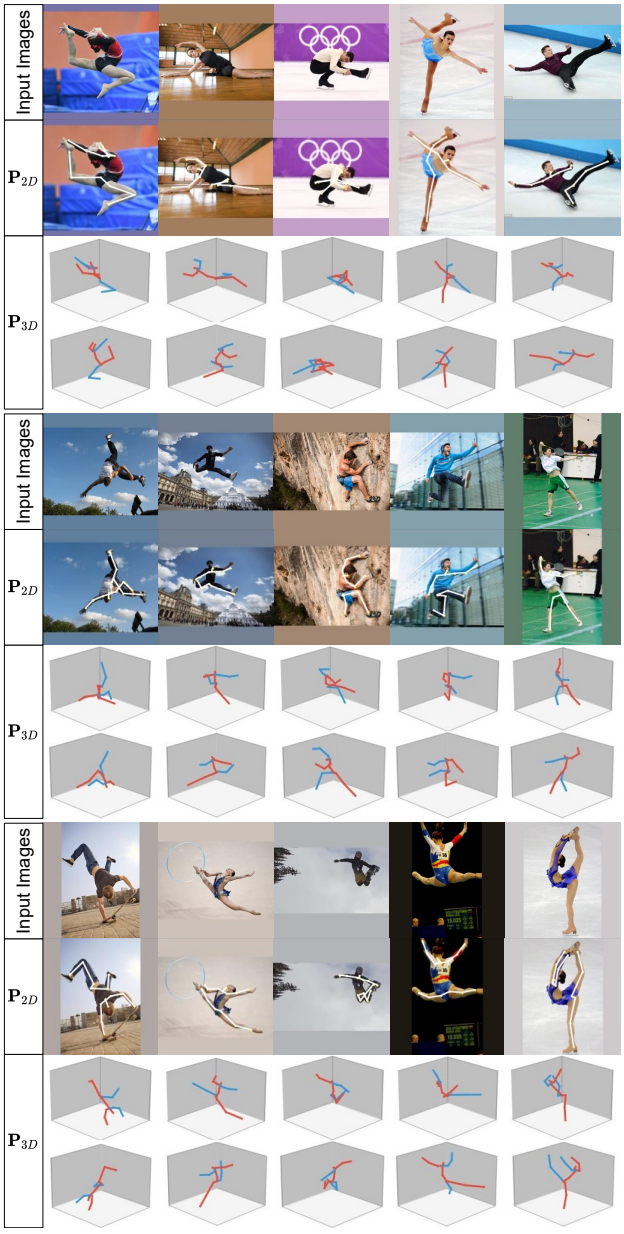

Figure 3. More human qualitative results on LSP [5].

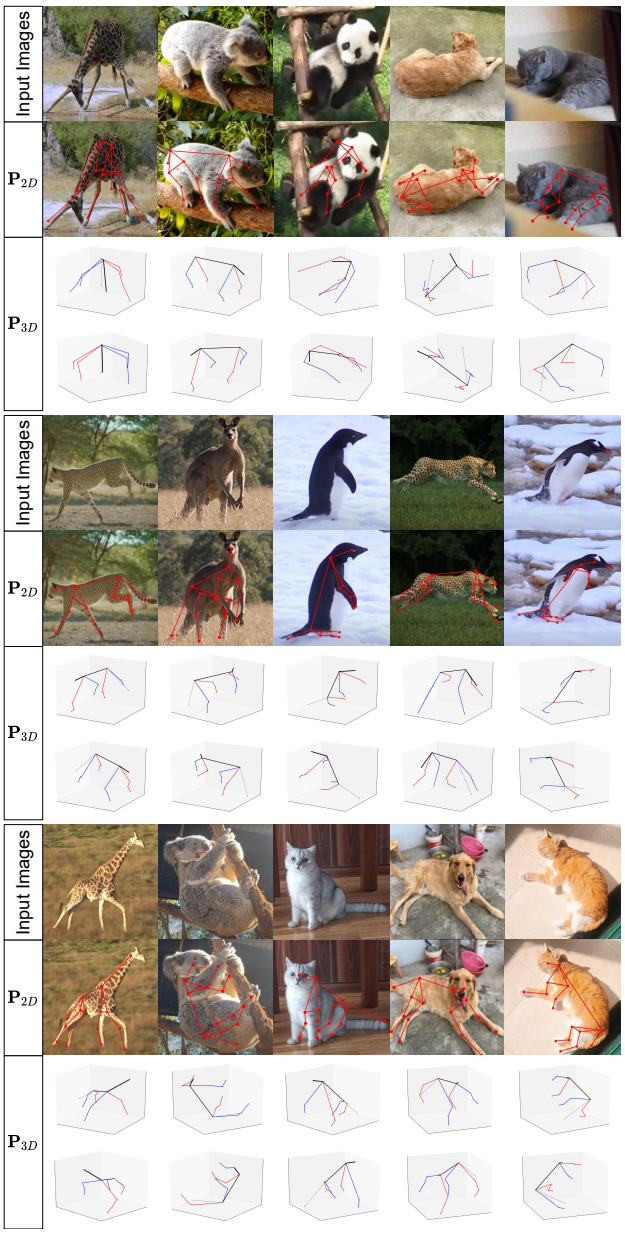

Figure 4. More animal qualitative results on self-collected data.