Monocular 3D Human Pose Estimation by Generation and Ordinal Ranking Supplementary Material

Saurabh Sharma[1] Pavan Teja Varigonda[2,3] Prashast Bindal[2] Abhishek Sharma[3] Arjun Jain[2,3]

[1]Max Planck Institute for Informatics Saarbrücken [2]Indian Institute of Technology Bombay [3]Axogyan AI Bangalore

Here we present the additional implementation details for MultiPoseNet and OrdinalNet, qualitative results on Human3.6[5] and results of other additional experiments.

1. Implementation Details

1.1. Architecture of MultiPoseNet

We follow [7] and employ simple, multilayer fully connected layers of dimensionality 1024 with Rectified Linear Units(ReLU) [2], Dropout [8], Batch-normalization [4], and Residual Connections [3] for our MultiPoseNet module.

The Encoder takes an input of size (16*2)(2D pose J_{2D}) + (17*3)(3D pose J_{3D}) and outputs the mean and covariance of $q(\hat{z}|J_{3D}, J_{2D})$, which are of size 256 each. The Encoder first processes the input to a size of 1024 using $FC(1024) - BN - ReLU - Dropout(0.5)$, followed by two $ResidualBlocks$, and finally applies $FC(512)$ to get the mean and covariance of the posterior. To sample $\hat{z}$ from the posterior we use the reparameterization trick from [6].

The Decoder transforms an input of size 256(latent code z) + 768(2D pose embedding) to an output of size 51(3D pose $\hat{J}_{3D}$). Analogous to the Encoder design, we use $FC(768) - BN - ReLU - Dropout(0.5)$ to get a 2D pose embedding from J_{2D}, which is then concatenated with the sampled latent code vector z and fed to two $ResidualBlocks$, and finally we use $FC(51)$ to get the output 3D pose $\hat{J}_{3D}$.

The architecture of each $ResidualBlock$ is $FC(1024) - BN - ReLU - Dropout(0.5) - FC(1024) - BN - ReLU - Dropout(0.5)$. Note the numbers in brackets indicate output dimensionality for FC layers and retention probability for Dropout layers. This architecture performed the best on the Human3.6 validation set in terms of average error and sample diversity, across a wide selection of hyperparameter choices for (1) the number of $ResidualBlocks$ and (2) size of the latent code z.

1.2. Ordinal Maps

We visualise the Ordinal Maps presented by our OrdinalNet in Fig.1.

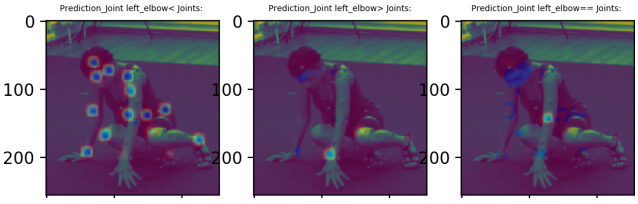

Figure 1: Predicted ordinal maps for the left elbow for test sample from Human3.6. The left wrist joint is closer to the camera than left elbow, and is predicted in $OM_{2,left\_elbow}$ map, while other joints greater in depth are predicted in $OM_{1,left\_elbow}$ map.

2. Additional Experiments

2.1. Qualitative Results

For qualitative analysis, we show the output of our model for a diverse set of poses on the Human3.6 test set in Fig.2. Note that here we synthesize 200 samples and use an Oracle to pick the best sample. The results indicate that 3D pose estimates from our model are quite good, across a wide range of poses.

2.2. Interpolation in Latent Space

To demonstrate that the manifold learned by the CVAE is smooth and semantically meaningful, we fix a 2D pose and interpolate between two randomly sampled noise vectors and show the generated samples in Fig. 3. The interpolation shows smooth variation in the generated 3D poses, while the 2D projection remains largely consistent. It follows that the CVAE has learnt a meaningful latent space which encodes only the depth variations in the joints. This validates our choice of the CVAE as a 2D-3D generative model that reduces the ambiguity in lifting from 2D-3D.

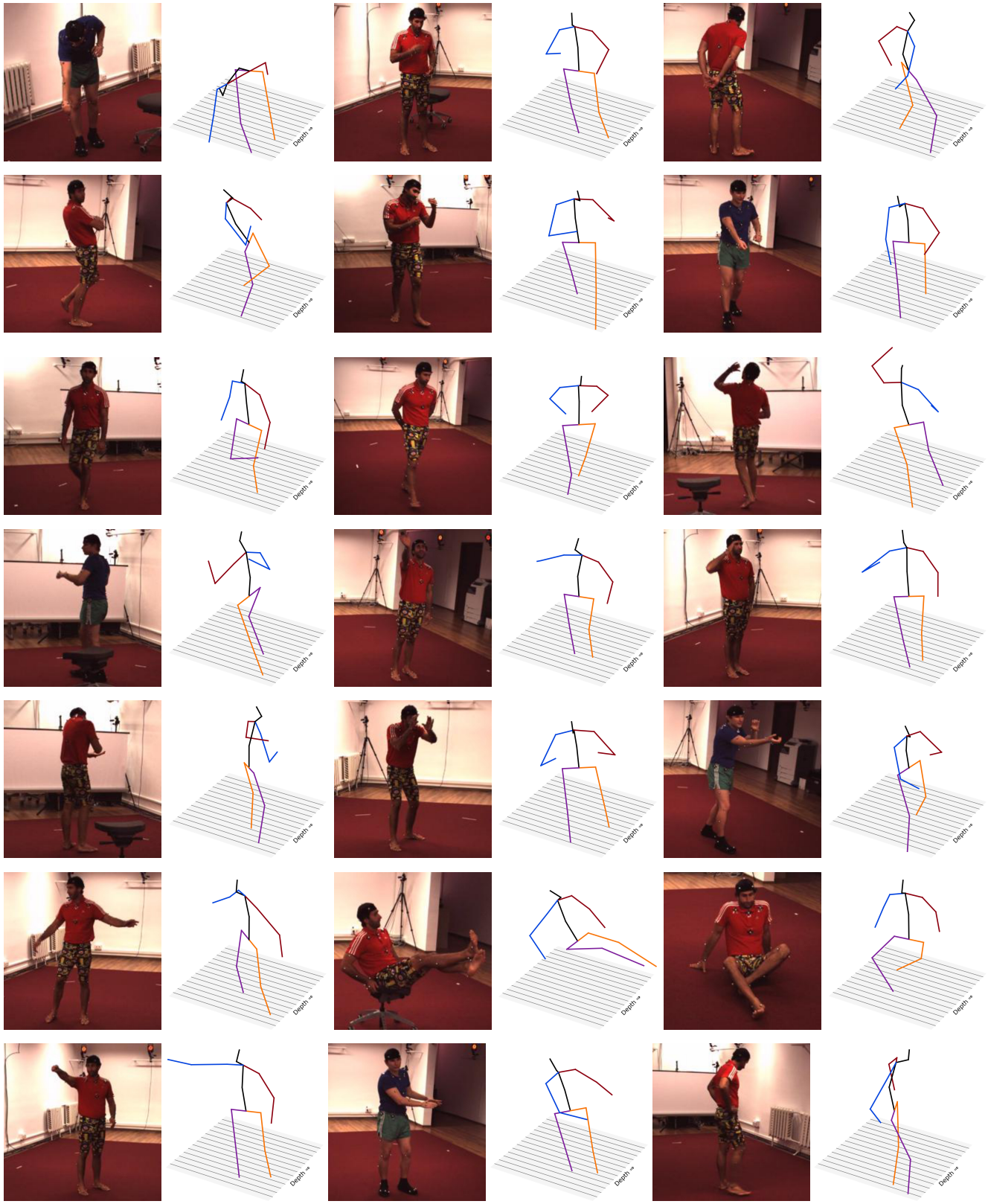

Figure 2: Qualitative results on the test set of Human3.6. The visualized 3D pose is the best sample chosen by the Oracle from a sample set of 200. Please note that for the poses, the azimuth is at an offset of 45°to the camera for ease of viewing.

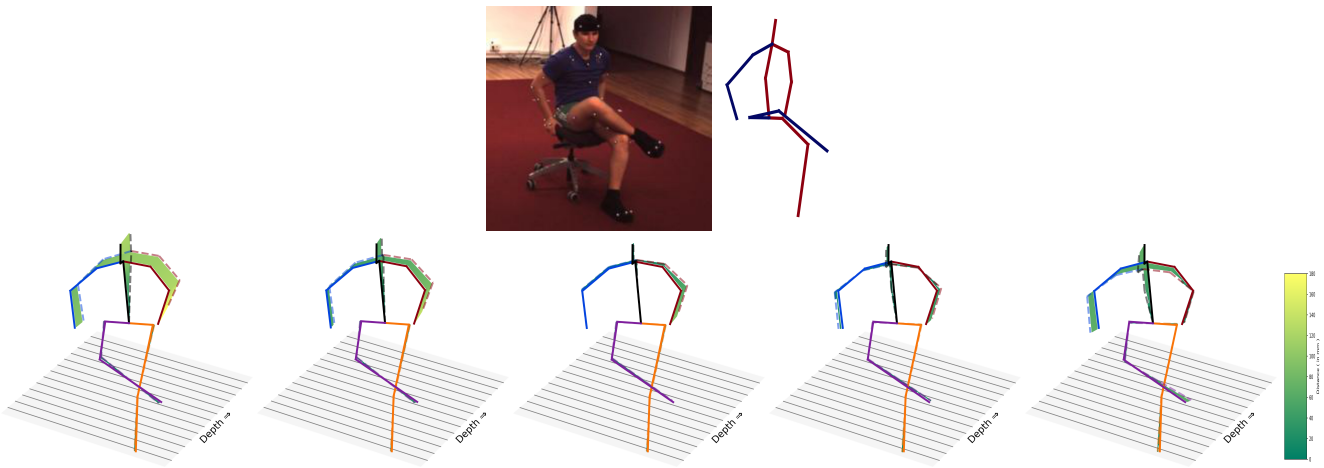

Figure 3: Latent space interpolations between two randomly sampled noise vectors z_1 and z_2 conditioned on the same 2D pose $\hat{J}_{2D}$. Above - Input Image and the detected 2D pose $\hat{J}_{2D}$. Below, from left to right - $Dec(z_1, \hat{J}_{2D})$, $Dec((3{*}z_1{+}z_2)/4, \hat{J}_{2D})$, $Dec((2 * z_1 + 2 * z_2)/4, \hat{J}_{2D})$, $Dec((z_1 + 3 * z_2)/4, \hat{J}_{2D})$, $Dec(z_2, \hat{J}_{2D})$. Mean pose is solid and sample is dashed, with displacement vector field in between.

2.3. Anatomical Consistency of Generated 3D Poses

A good sampling mechanism for 3D pose estimation should generate anthropomorphically valid samples with a high probability. To quantitatively assess this point, we use the PosePrior model from [1], that uses pose-conditioned joint-angle constraints to classify a 3D pose as valid/invalid. We use their publicly available code to compute the percentage of samples generated by MultiPoseNet that are anthropomorphically valid. We rank the generated 3D candidate set using *OrdinalScore* with both the ground-truth and predicted ordinals, and plot the valid sample percentage in the interval [1,x] as x varies from 1 to 100. This is depicted in Fig. 4.

The plot shows that for the interval [1,100], 90% of the samples are valid, confirming our intuition that the CVAE generates valid 3D pose candidates. It also demonstrates that the ordinal score and anthropomorphic validity are correlated.

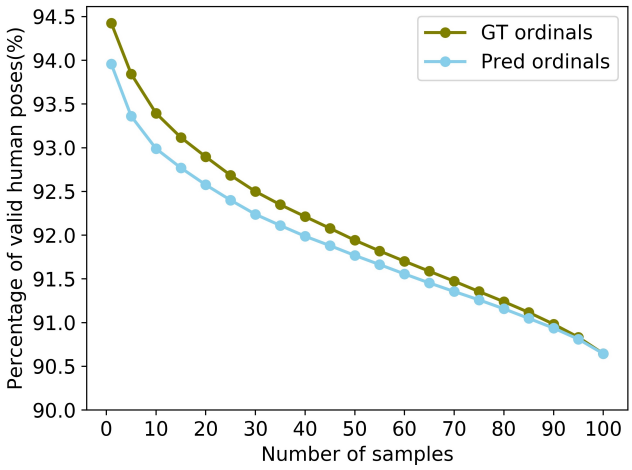

Figure 4: Cumulative percentage of anatomically valid samples according to PosePrior[1] for 100 generated samples.

References

[1] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *CVPR*, 2015. 3

[2] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *AISTATS*, 2011. 1

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

[4] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 1

[5] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014. 1

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2013. 1

[7] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 1

[8] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014. 1