Ordinal Depth Supervision for 3D Human Pose Estimation Supplementary material

Georgios Pavlakos[1], Xiaowei Zhou[2], Kostas Daniilidis[1]
[1] University of Pennsylvania [2] State Key Lab of CAD&CG, Zhejiang University

This supplementary material provides additional details that were not included in the main manuscript due to space constraints. Sections 1 and 2 present some additional qualitative evaluations of our approach. Section 3 includes a detailed description of the different architectures employed in our experiments. Section 4 provides more details about the training procedure we followed. Section 5 clarifies details of our quantitative evaluation. Finally, Section 6 focuses on extensive qualitative evaluation of our approach to complement the quantitative evaluation of the main manuscript.

1. Reconstruction component

To assess the effectiveness of ordinal depth relations for 3D human pose estimation, we evaluate our reconstruction component using a) only 2D keypoints as input (i.e., [4]) and b) concatenating 2D keypoints and the ordinal depth relations of the joints. Table 1 presents the results for the two versions, where the architecture is the same and ground truth 2D keypoints and ordinal relations are used for training and testing. The decrease of the average 3D error is indeed expected when we add ordinal depth information. However, the level of improvement is quite significant, achieving relative error reduction greater than 30% when we include the ordinal depth relations to the input. This comparison provides additional evidence that ordinal depth relations encode substantial information and can indeed boost the performance of approaches that employ them (e.g. discriminative methods). We clarify here that our implementation of the reconstruction prior for the comparison with the state-of-the-art uses the predicted depth values z_n from the initial ConvNet as input, instead of the ordinal relations, however, the ordinal depth relations can be a good indication of the information that is encoded by the predicted z_n values.

Input	Avg Error
2D keypoints (GT) [4]	45.5
2D keypoints + Ordinal relations (GT)	**31.6**

Table 1: Evaluation on Human3.6M when a reconstruction component is employed using a) only 2D keypoints as input, or b) combining 2D keypoints with the ordinal depth relations of the joints. All inputs are the ground truth values. The numbers are mean per joint errors (mm). The number of the first row is taken from the respective paper. The addition of ordinal depth leads to a substantial relative error reduction that exceeds 30%.

	Human agreement rate$^{\neq}$ (%)
Zhou *et al*. [12]	71.72
Bogo *et al*. [1]	73.06
Ours	**85.86**

Table 2: Human agreement rates on the ordinal annotations of the LSP test set, for pairs annotated with an $\neq$ depth relation. Our image-based ConvNet outperforms state-of-the-art reconstruction approaches which employ only the detected 2D joints and rely on 3D body shape priors to resolve reconstruction ambiguities and ignore additional image evidence.

2. LSP+MPII Ordinal

Although LSP+MPII Ordinal is not appropriate for mm level evaluation, we can still use it for empirical comparison by considering the agreement of the predicted ordinal depth relations with the relative annotations provided by humans. To further stress the importance of image-based information for 3D reconstruction, which is explicitly leveraged by our method, we compare with two state-of-the-art approaches that reconstruction 3D given only 2D correspondences, Zhou *et al*. [12] and Bogo *et al*. [1]. Similar to [1], the input 2D joints for these methods are localized using the DeepCut 2D pose detector [9]. The human agreement rates for the test set of LSP dataset are presented in Table 2, where the "neck" joint has been ignored, since the method of [1] doesn't produce an estimate for it. As we can see, our approach clearly outperforms both reconstruction approaches on this metric. This indicates that by relying only on 2D joint locations, reconstruction approaches are under-using the available information and could potentially benefit from the ordinal depth output of our ConvNet to produce more convincing reconstructions.

3. Architecture

Our empirical evaluation focused on the benefit of the supervision with ordinal depth, regardless of the particular

representation (e.g. coordinate or volume regression), or architecture. For completeness, here we provide more details regarding the exact network architectures we used in our experiments. As we described in Section 4.2 of the main manuscript, the main building block for the majority of the experiments is a hourglass module which follows the design of [7]. The exact architecture of the hourglass module is presented in Figure 1. For the experiments of Table 1 of the main manuscript, we use the same component, and we only change the output, i.e., we add a fully connected layer at the end with N outputs for depth prediction, we add a fully connected layer at the end with $3N$ outputs for coordinate regression, we add a 1×1 convolutional layer to produce $N \times 64$ channels for the volumetric output. For the experiment with the two hourglasses we adopt the coarse-to-fine scheme of [8], to be compatible with them. We give more details about this architecture in Figure 2. For the rest of the experiments (Tables 2-7 of the main manuscript), this stacked architecture is used as the base ConvNet, while the reconstruction component is attached at the end (following Figure 3b of the main manuscript). This component is a simple multilayer perceptron similar to [4], while the exact architecture is described in Figure 3.

4. Training details

For the experiments of Table 1 of the main manuscript, comparing weak ordinal supervision with full 3D supervision, our weakly supervised versions are trained using only the weak ordinal losses for each of the three cases (depth prediction detailed in Section 3.1, coordinate regression detailed in Section 3.2, and volumetric regression detailed in Section 3.3). For the rest of the experiments (Tables 2-7 of the main manuscript), where we compare with the state-of-the-art, we use a mixed training strategy (described in Section 4.2 of the main manuscript). For the images coming from the 3D dataset (Human3.6M or HumanEva-I), we use full supervision for the output volume, while for the images coming from LSP+MPII Ordinal, we calculate the loss based on our weak ordinal loss (decomposed for 2D keypoints and ordinal depth). The batches for training are drawn randomly and can contain images from both sources. Depending on the source of the image, the loss is calculated accordingly. Regarding the reconstruction component, it is trained independently using only MoCap data, as it is described in Section 4.2 of the main manuscript. The whole system combining the ConvNet and the reconstruction module can also be refined end-to-end, but we found the effect of this refinement to be marginal if both of the two components are well trained.

5. Evaluation details

The results of Table 1 of the main manuscript are the only case where the architectures do not include the reconstruction component at the end, since our focus is to compare the type of supervision (weak with ordinal relations versus full supervision). As a result, the ordinal depth predictions are

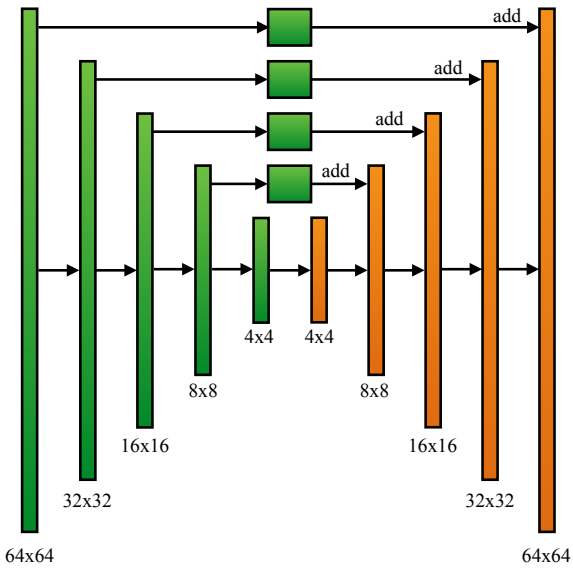

Figure 1: Detailed architecture for the hourglass module. The columns of the figure indicate residual modules which enclose the convolutional layers. The kernels for the convolutions have size 3×3. The activation is ReLU, while batch normalization is also used. For the encoding part, max pooling (2×2) is used for the subsampling, while for the decoding part, nearest neighbor upsampling is used for the upsampling. The numbers below each column indicate the resolution of the feature map at every stage of the hourglass. The green color of the encoding modules indicate three consecutive residual modules, while the orange color of the decoding part indicates one residual module. The skip connections include also residual modules, and their output is fused with the feature maps of the main pathway after the upsampling, using element-wise addition. The number of channels is constant across the hourglass, and equal to 256.

not "filtered" through the reconstruction component, which helps producing a coherent 3D pose. In that case, to get metric predictions of the depth with respect to the root joint, we simply rescale the estimates of the ConvNet. In more detail, after the training has finished, we apply our model on a small set of the training data. Denoting with $\hat{d}$ all the predicted depth values, and with d the corresponding ground truth metric depths, we compute a scaling factor:

$$\alpha = \frac{\max d - \min d}{\max \hat{d} - \min \hat{d}}. \quad (1)$$

We use this scaling factor α to multiply the depth predictions from our network and get depth in metric scale. Using the predicted 2D location in pixel coordinates and assuming the camera intrinsics are known, we can reconstruct the 3D pose and be comparable with [8], that we use as a baseline here.

In Table 4 of the main manuscript we did not include in the comparison the results from the work of Sun *et al.* [11]

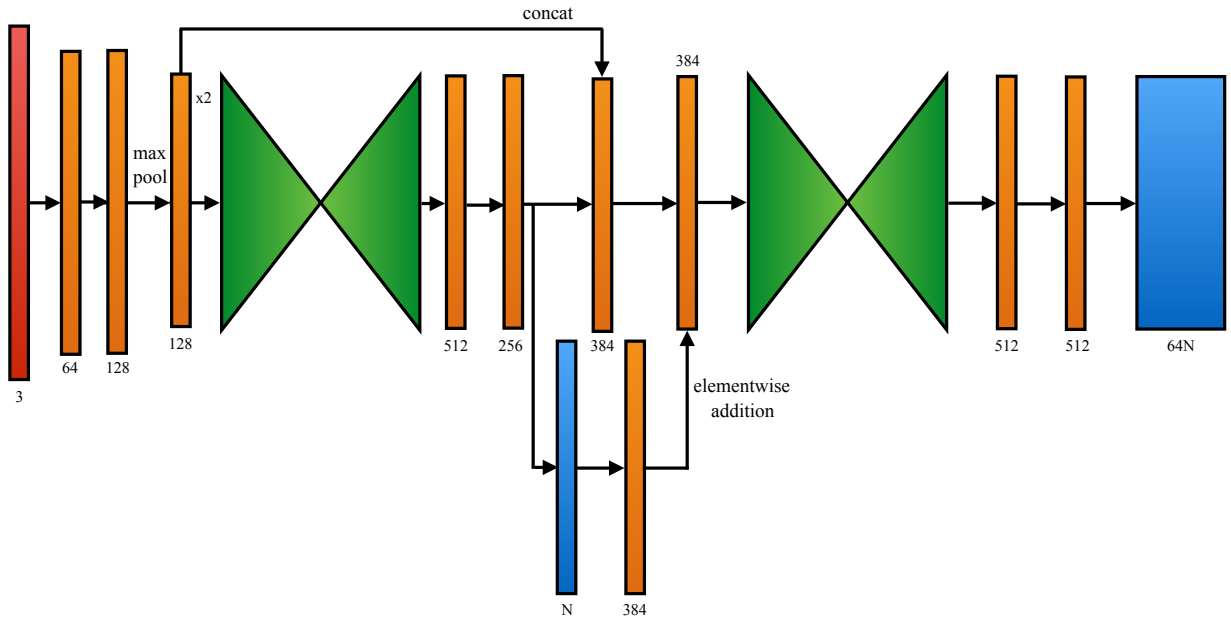

Figure 2: The detailed architecture used for the experiments with the two hourglasses. It has a coarse-to-fine scheme, where the output of the first hourglass is effectively 2D heatmaps, while the output of the second hourglass is volumetric. The green hourglasses have the design detailed in Figure 1, the orange columns are convolutional layers, the red column indicates the input image, while the blue columns correspond to the heatmaps (in 2D form as intermediate supervision for the first hourglass, and in 3D form as the final output for the second hourglass). The numbers on the bottom of each column indicate the number of channels for the feature maps. The convolutional layers are implemented as residual modules with 3×3 kernels. The only exceptions are the first layer which is a 7×7 convolution, and the two layers after each hourglass, as well as the layers that produce and post-process the outputs, that implement 1×1 convolutions. The first layer uses stride equal to two, decreasing the resolution from 256×256 of the original image to 128×128. After the second module, a max pooling decreases further the spatial resolution to 64×64, which remains constant until the end of the network (excluding the interior of the hourglasses).

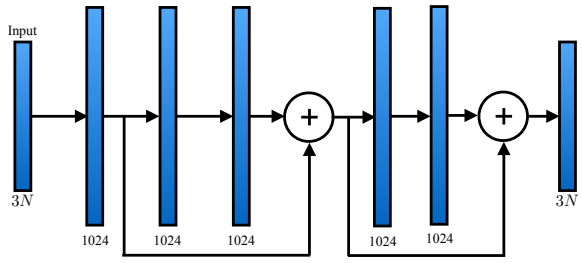

Figure 3: Detailed architecture for the reconstruction component. The input is a vector of size $3N$ (estimated 2D joint coordinates normalized in $[-1, 1]$ and predicted ordinal depth for N joints). Each column represents a fully connected layer (linear layer), except for the first column that corresponds to the input. The sizes of the fully connected layers are 1024, except for the last layer that is of size $3N$. After each fully connected layer of size 1024, we use batch normalization, ReLU, and Dropout. The final output is a vector of size $3N$, corresponding to the 3D pose coordinates.

(reported average error of 59.1mm), since the authors confirmed to us through personal communication that they use the depth of the root joint to reconstruct the full 3D pose. This makes the results not directly comparable with the rest of the approaches on this Table. Under this setting, we achieve even better numbers (53.0mm average error), but this is comparable only with [11].

6. Qualitative evaluation

The main manuscript presented extensive quantitative evaluation of our approach. Here we provide some additional qualitative results on the various datasets. Figure 4 collects successful reconstructions from Human3.6M [2], using our state-of-the-art model from Table 2, 4 and 5 of the main manuscript. Since we are using a mixed training strategy for this dataset, including data from Human3.6M itself, the failures are minimal (Figure 5). Most of them can be attributed to erroneous 2D localization of the joints, challenging 3D poses, either because of heavy self-occlusions, or because of rarity in the training set (compared to the dominant standing poses) while left-right flipping can also be a (rare) source of error. Similar qualitative results are provided also for HumanEva-I [10] in Figure 6, where the reconstructions are even more accurate, because the same users appear in both the training and test set.

Besides these datasets, we also provide qualitative results for MPI-INF-3DHP [5, 6], where the domain shift is significant compared to Human3.6M. Typical reconstructions for this benchmark are collected in Figure 7. In general, the backgrounds are different from Human3.6M, the subjects are acting with more freedom, no markers are attached to them, and outdoor captures are also included. The dataset cannot be considered truly in-the-wild, but we emphasize that the model is reliable even if no data from this benchmark has been used for training (as we detail in the main manuscript, we use only data from Human3.6M and LSP+MPII Ordinal). To underline the importance of our ordinal annotations for the proper generalization of the model, we compare the qualitative results of three different models (following the quantitative evaluation of the main manuscript in Table 3). These three models are trained using: a) only Human3.6M data for training b) Human3.6M data and LSP+MPII images with supervision from 2D keypoints only, and c) Human3.6M data and LSP+MPII Ordinal images with supervision from 2D keypoints *and* ordinal depth annotations. The qualitative results of this comparison are presented in Table 8, with the first three columns corresponding to the first model. the next three columns to the second model, and the last three columns to the third model. Unsurprisingly, the Human3.6M model performs very unreliably, because of heavy overfitting on Human3.6M data. Adding in-the-wild images with only 2D supervision improves significantly the 2D aspect of the detector, but the depth prediction is still mediocre. It is crucial to incorporate the ordinal depth annotations in the training procedure to get a model that achieves reliable prediction both for 2D keypoint locations and for depths of the joints.

Finally, for images that are considered in-the-wild, we present qualitative results on the test set of the LSP dataset [3] in Figure 9. The model in this case is similar to the one we used earlier, with the exception that we used no images from the LSP test set for training. Since in-the-wild images with 2D keypoint and ordinal depth annotations have been incorporated in the training (data from LSP+MPII Ordinal), the 3D reconstructions are again reasonable.

References

[1] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 1

[2] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014. 3, 5

[3] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 4, 9

[4] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 1, 2

[5] D. Mehta, H. Rhodin, D. Casas, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, 2017. 4, 6, 7, 8

[6] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. VNect: Real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics*, 36, 2017. 4, 6, 7, 8

[7] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[8] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017. 2

[9] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 1

[10] L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4–27, 2010. 3, 5

[11] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *ICCV*, 2017. 2, 3

[12] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis. Sparse representation for 3D shape estimation: A convex relaxation approach. *PAMI*, 2016. 1

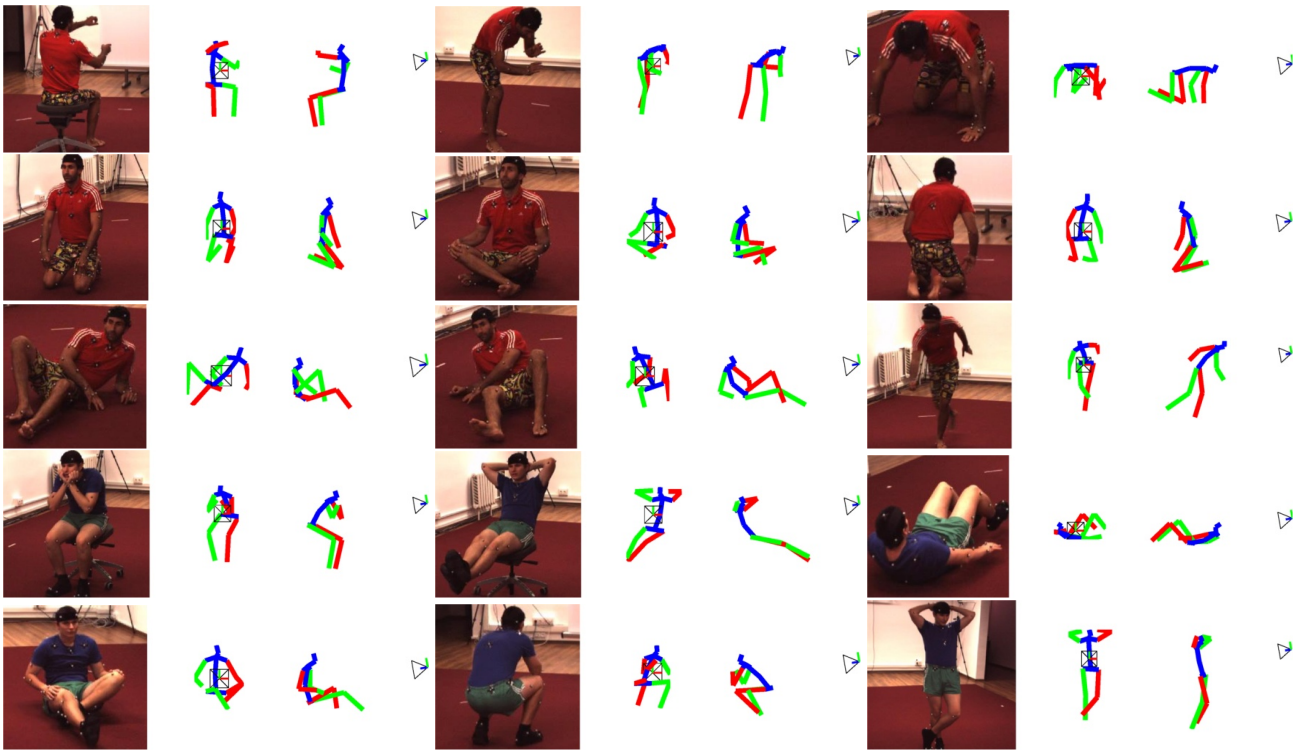

Figure 4: Successful reconstructions on Human3.6M [2]. For each example, we present the test image, and the predicted 3D pose from the original view, and a novel view.

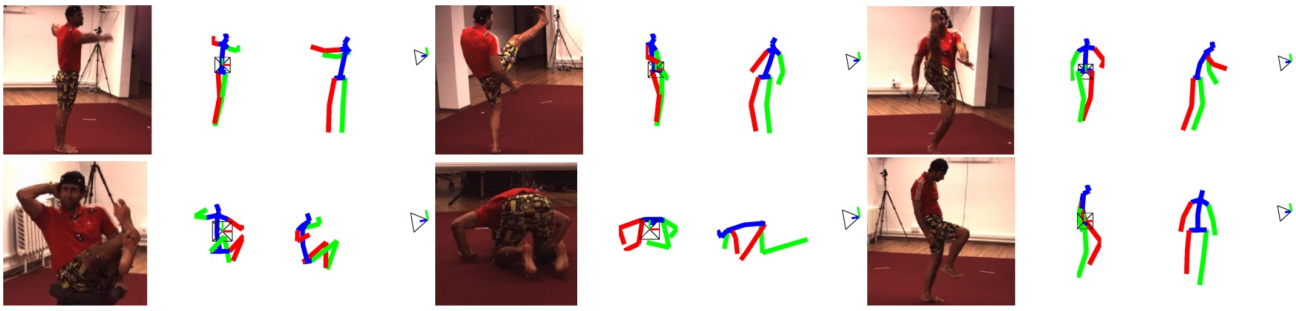

Figure 5: Erroneous reconstructions on Human3.6M [2]. For each example, we present the test image, and the predicted 3D pose from the original view, and a novel view. The main failures can be attributed to joints that are not correctly localized on the 2D image, poses that are very challenging, because of heavy self-occlusions, or because they are rare (compared to the dominant standing poses of the training set), or to some rare cases with left-right flipping.

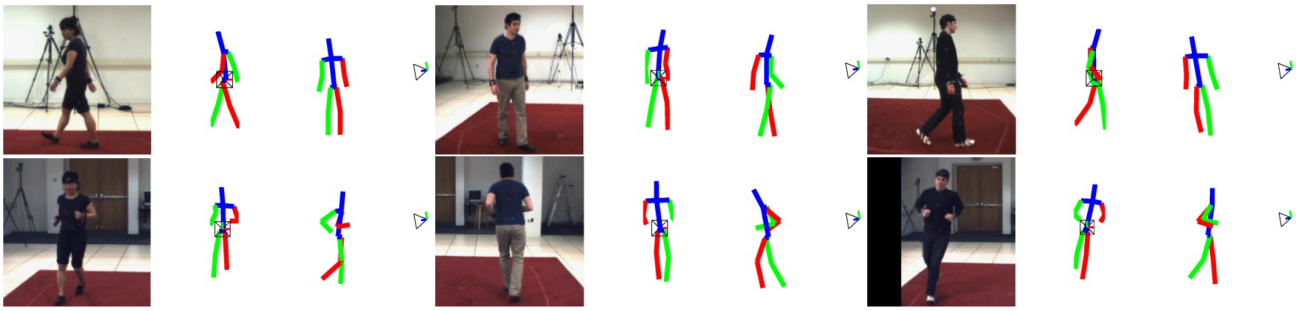

Figure 6: Qualitative results on HumanEva-I [10]. For each example, we present the test image, and the predicted 3D pose from the original view, and a novel view.

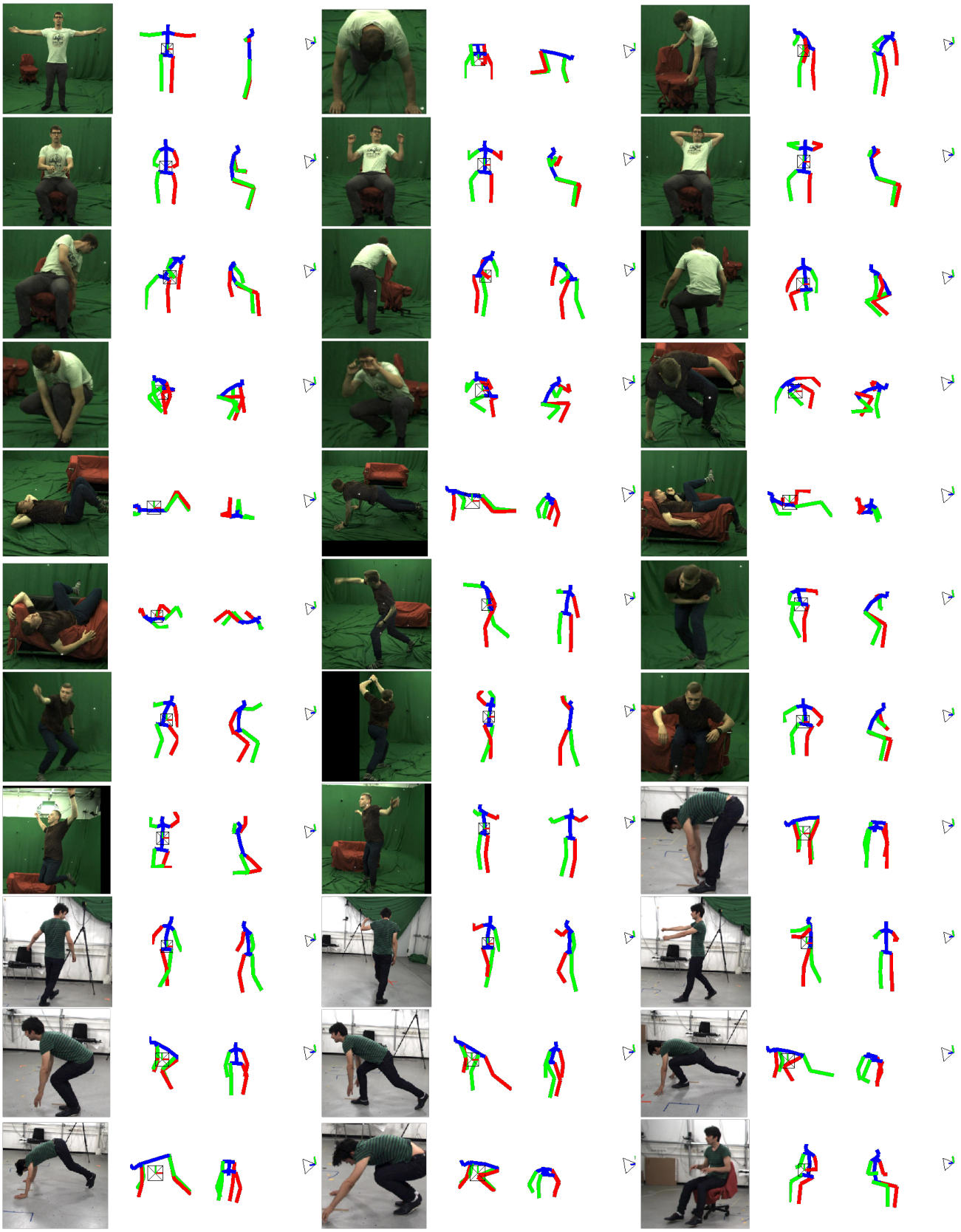

Figure 7: Qualitative results on MPI-INF-3DHP [5, 6]. For each example, we present the test image, and the predicted 3D pose from the original view, and a novel view. We emphasize that our model has not been trained on this dataset.

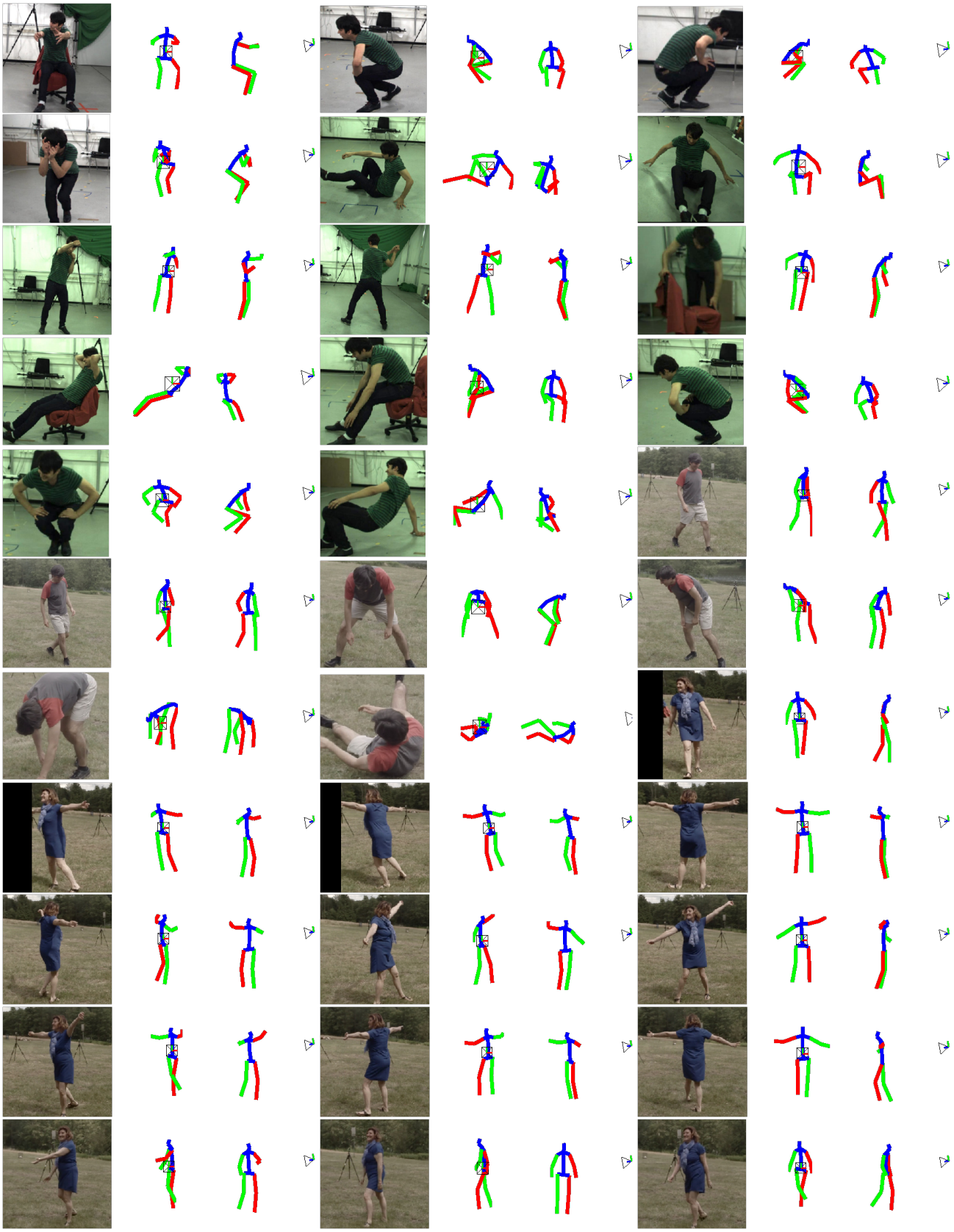

Figure 7 (cont.): Qualitative results on MPI-INF-3DHP [5, 6]. For each example, we present the test image, and the predicted 3D pose from the original view, and a novel view. We emphasize that our model has not been trained on this dataset.

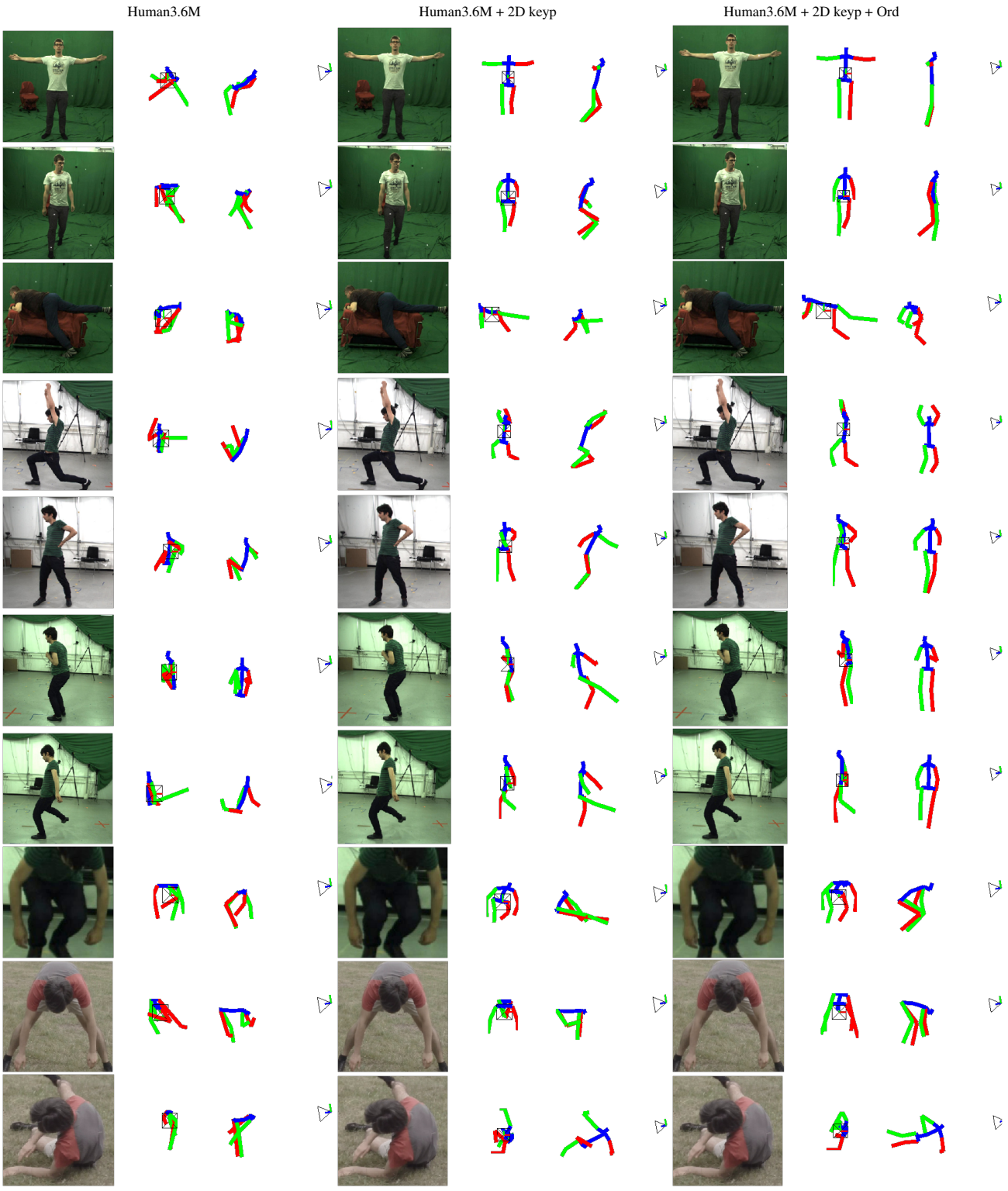

Figure 8: Qualitative evaluation on MPI-INF-3DHP [5, 6], demonstrating the importance of ordinal depth annotations for proper generalization. Each row corresponds to one image example, where we present the test image and the predicted 3D pose for the three different models discussed in Table 3 of the main manuscript. The first three columns refer to the first model (training with Human3.6M data), the three middle columns to the second model (training with Human3.6M data and in-the-wild images with 2D keypoint annotations), and the last three columns to the third model (training with Human3.6M data and in-the-wild images with 2D keypoint and ordinal depth annotations).

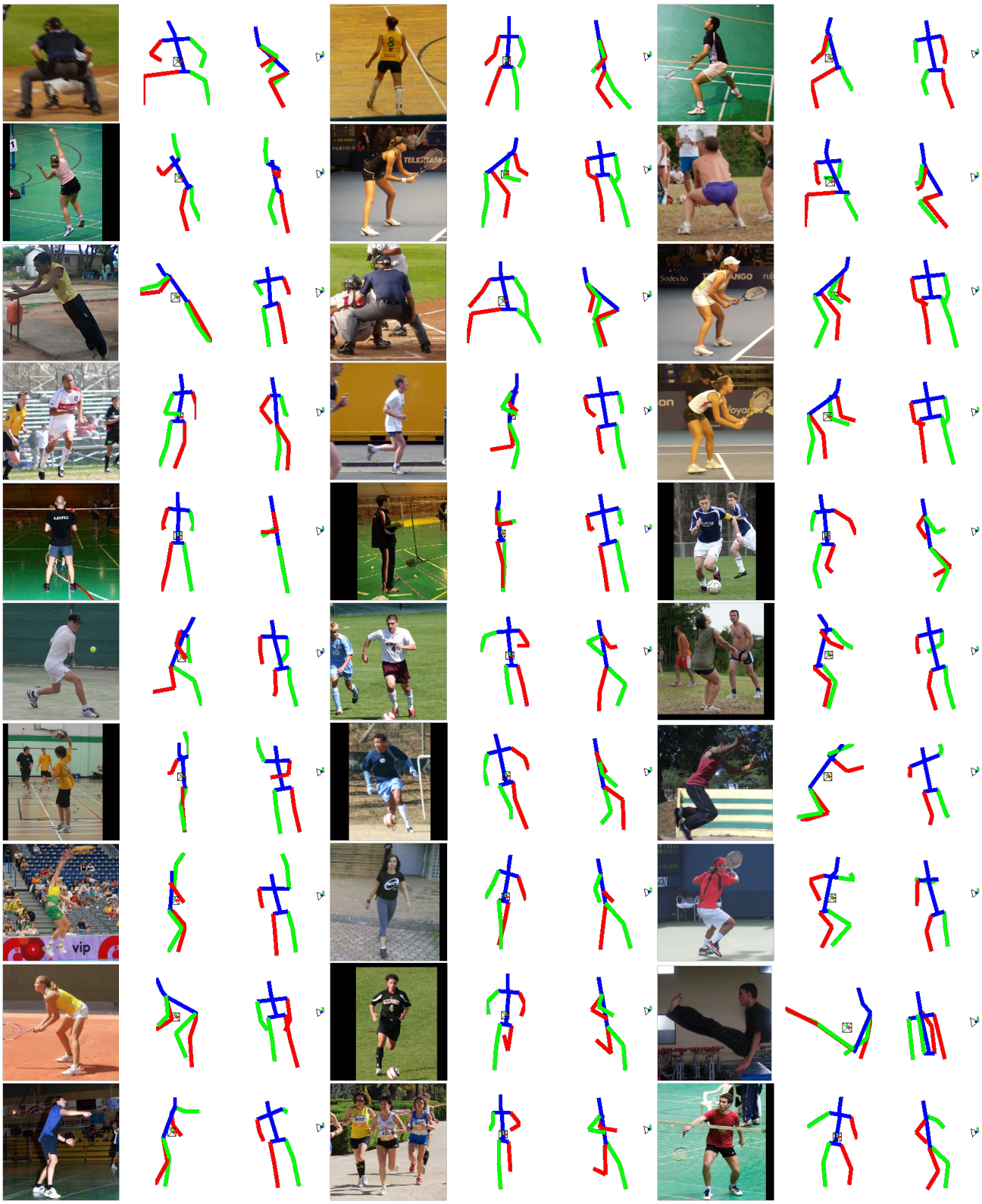

Figure 9: Successful reconstructions on the test set of LSP [3]. For each example, we present the test image, and the predicted 3D pose from the original view, and a novel view.