Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose Supplementary Material

Georgios Pavlakos[1], Xiaowei Zhou[1], Konstantinos G. Derpanis[2], Kostas Daniilidis[1]
[1] University of Pennsylvania [2] Ryerson University

In this supplementary, we provide material that could not be included in the main manuscript due to space constraints. First, Section 1 provides additional details for the exact representation of the volumetric space and the way we can obtain metric pose estimates from the voxelized estimates. Section 2 presents full results on Human3.6M using the reconstruction error for evaluation. Section 3 further examines the decoupled architecture in the case the groundtruth 2D joints are provided as input, while Section 4 focuses on quantitative evaluation on the MPII Human Pose dataset using our volumetric representation. Finally, Section 5 presents additional quantitative results and provides insights about the failure cases or our approach.

1. Details of volumetric representation

As detailed in the main manuscript, in our definition of the volumetric space, the x-y dimensions correspond to pixel coordinates, while the z dimension corresponds to metric depth with respect to the root joint. Let us assume that the ConvNet predictions for joint i are $[x_i, y_i, z_i]$. Given the depth of the root joint, Z_{root} and the camera calibration matrix, K, the metric position S_i of joint i is computed as:

$$S_i = (Z_{root} + z_i)K^{-1}[x_i, y_i, 1]^\top. \tag{1}$$

This means that the reconstruction is recovered up to the depth of the root joint, Z_{root}. If this information is unknown, we propose to estimate it based on the expected size of the skeleton by solving:

$$\min_{Z_{root}} \sum_{(i,j)\in\mathcal{S}} \left(\|S_i - S_j\| - L_{ij}\right)^2, \tag{2}$$

where $\mathcal{S}$ is the set of pairs of skeleton joints i, j that are connected and L_{ij} is the corresponding limb length. Effectively, problem 2 minimizes the discrepancy between the limb lengths of the predicted 3D pose and the desired values L_{ij}. This problem is convex and can be solved very efficiently as there is only one variable.

Based on the available information during the reconstruction, we evaluate the following scenarios on Human3.6M:

- the depth of the root node is provided (depth root);
- L_{ij} are given by the subject ground truth skeleton (personal skeleton);
- L_{ij} are given by their mean values across training subjects (universal skeleton).

The results for the different scenarios are presented in Table 1. The performance difference is small and is justified by the additional information provided in different scenarios, e.g., the benefit of using a subject-specific versus a universal skeleton. Other single view approaches require knowledge of the root joint location (e.g., [2]). This is also a common assumption among motion capture methods, e.g., [3], since the root (pelvic) joint is usually easy to track [4]. We reiterate that for the comparison with the state-of-the-art we only compute an estimate of the depth of the root joint, based on the univeral skeleton of each dataset, using the procedure we described above.

2. Reconstruction error results on Human3.6M

In Table 5 of the main manuscript we presented quantitative results for Human3.6M using the reconstruction error for evaluation. Due to space constraints only the average error was included. We extend these results here, by presenting our performance for all actions in Table 2.

3. Decoupled with 2D groundtruth as input

In Section 4.4 of the main manuscript, the decoupled architecture was applied directly on the image. An alternative setting is to use the groundtruth 2D joints as input to the component that is responsible for 2D-to-3D reconstruction. This will give us an estimate of the ideally optimal performance we can expect from the decoupled architecture. Table 3 presents a comparison between this version of the decoupled architecture against the typical approach of using as input a network's predictions for the 2D joint locations. As expected, using groundtruth 2D locations is beneficial compared to the (possibly erroneous) predictions from another ConvNet. Interestingly, the error reduction is not significant. This indicates that the most challenging part of the prediction is the 3D reconstruction, rather than 2D localization. In fact, prediction is comparable to our Coarse-to-Fine architecture trained end-to-end (69.77mm), demonstrating the importance of using image-based evidence for the 3D prediction, instead of relying exclusively on 2D joint locations.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Universal skeleton	67.38	71.95	66.70	69.07	71.95	76.97	65.03	68.30
Personal skeleton	60.92	67.10	61.85	62.85	67.53	72.27	58.97	64.37
Depth root	59.32	64.87	59.48	61.25	65.12	69.02	57.06	60.14
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Average
Universal skeleton	83.66	96.51	71.74	65.83	74.89	59.11	63.24	71.90
Personal skeleton	79.84	92.88	67.03	60.95	70.97	54.05	57.65	67.07
Depth root	75.14	91.89	64.51	59.55	66.81	53.67	56.79	64.76

Table 1: Quantitative comparison of our approach on Human3.6M under various evaluation scenarios. The numbers are the average 3D joint errors in mm.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Ours	47.54	50.52	48.30	49.31	50.74	55.22	46.10	48.00
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Average
Ours	61.09	78.07	51.05	48.31	52.85	41.53	46.42	51.88

Table 2: Quantitative results of our approach on Human3.6M dataset. The numbers are reconstruction errors in mm.

2D Locations	Network Predictions	Groundtruth
Mean 3D error	78.10mm	67.87mm

Table 3: Comparison of the decoupled architecture on Human3.6M, where 2D joints are localized by a ConvNet or their groundtruth location is used as input (extends Table 3 of the main manuscript).

4. Quantitative evaluation on MPII

Since MPII does not provide 3D pose groundtruth, we focus on qualitative evaluation by presenting compelling 3D reconstructions from in-the-wild images. Nevertheless, an interesting quantitative experiment is to evaluate whether the 2D localization accuracy of the decoupled architecture is comparable to a ConvNet that has been trained explicitly for the 2D localization task. For a fair comparison, we use our decoupled architecture and we evaluate the localization accuracy of the 2D heatmaps (output of the network component trained explicitly for the 2D task), and the 3D heatmaps (output of the 2D-to-3D reconstruction component using the 2D heatmaps as input and producing 3D heatmaps). The results for the MPII validation set are presented in Table 4.

Interestingly, the performance of both outputs is comparable. 3D heatmaps obtain similar 2D localization accuracy with 2D heatmaps, with the advantage of providing the 3D reconstruction as well. Effectively, the 2D-to-3D reconstruction component acts as a regularization on the 2D heatmaps, producing compelling 3D predictions given the initial 2D estimates. The few failures that degrade 2D localization performance can be attributed to challenging poses which were not available in the 3D pose examples that the 2D-to-3D reconstruction component was trained on. In these challenging cases, the inference of the 2D-to-3D reconstruction component produces an alternative 3D prediction which might deteriorate localization for some joints.

5. Additional qualitative evaluation

We provide additional qualitative results using the proposed volumetric representation on a variety of examples. The included images are from MPII (Fig. 1, 2, 3 and 4), Human3.6M (Fig. 5), HumanEva-I (Fig. 6), and KTH Football II (Fig. 7). Finally, Figure 8 illustrates a qualitative comparison on Human3.6M between the "Decoupled" architecture and our proposed "Coarse-to-Fine" architecture (as described in Section 3.3 and evaluated quantitatively in Section 4.4 of the main manuscript).

Besides presenting more examples, we elaborate here on the failure cases, some of which are presented in Figure 4. For datasets like Human3.6M, or HumanEva-I where 3D groundtruth is available for end-to-end training, huge errors are rare. They usually happen in cases of severe self-occlusion (e.g., actions like Sitting Down) or concern the violation of structural constraints (e.g., inconsistent length of symmetric limbs, joint angles exceeding angle limits). Since the network learns the structure of the human body implicitly from data, and it is not explicitly penalized for structural violations, these errors might be more common compared to model-based approaches (e.g., [5, 1]. On the other hand for in-the-wild images, the errors are more frequent. Novel viewpoints, extreme 3D poses that are not present in our 3D training examples and errors in the initial 2D localization can lead the final predictions astray. Effectively, because of the constraint of using only 2D joint locations for 3D reconstruction, and the use of 3D data from a different domain (here MoCap data with less variability than in-the-wild poses), we expect the final predictions to be less accurate than the controlled cases of Human3.6M or HumanEva-I.

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. 2

	Ankles	Knees	Hips	Wrists	Elbows	Shoulders	Head	PCKh
2D heatmaps	78.53	81.16	82.98	81.60	86.93	93.22	96.20	86.24
3D heatmaps	75.41	79.31	81.29	80.42	85.73	93.05	95.86	84.96

Table 4: Localization accuracy on the MPII validation set using the standard PCKh metric for each joint and overall. The 2D output of a network which has been trained explicitly for this task ("2D heatmaps") is compared with the output of our 2D-to-3D reconstruction component ("3D heatmaps") where the same 2D heatmaps are given as input. The addition of 2D-to-3D inference only slightly degrades 2D localization, while also producing a 3D prediction at the same time.

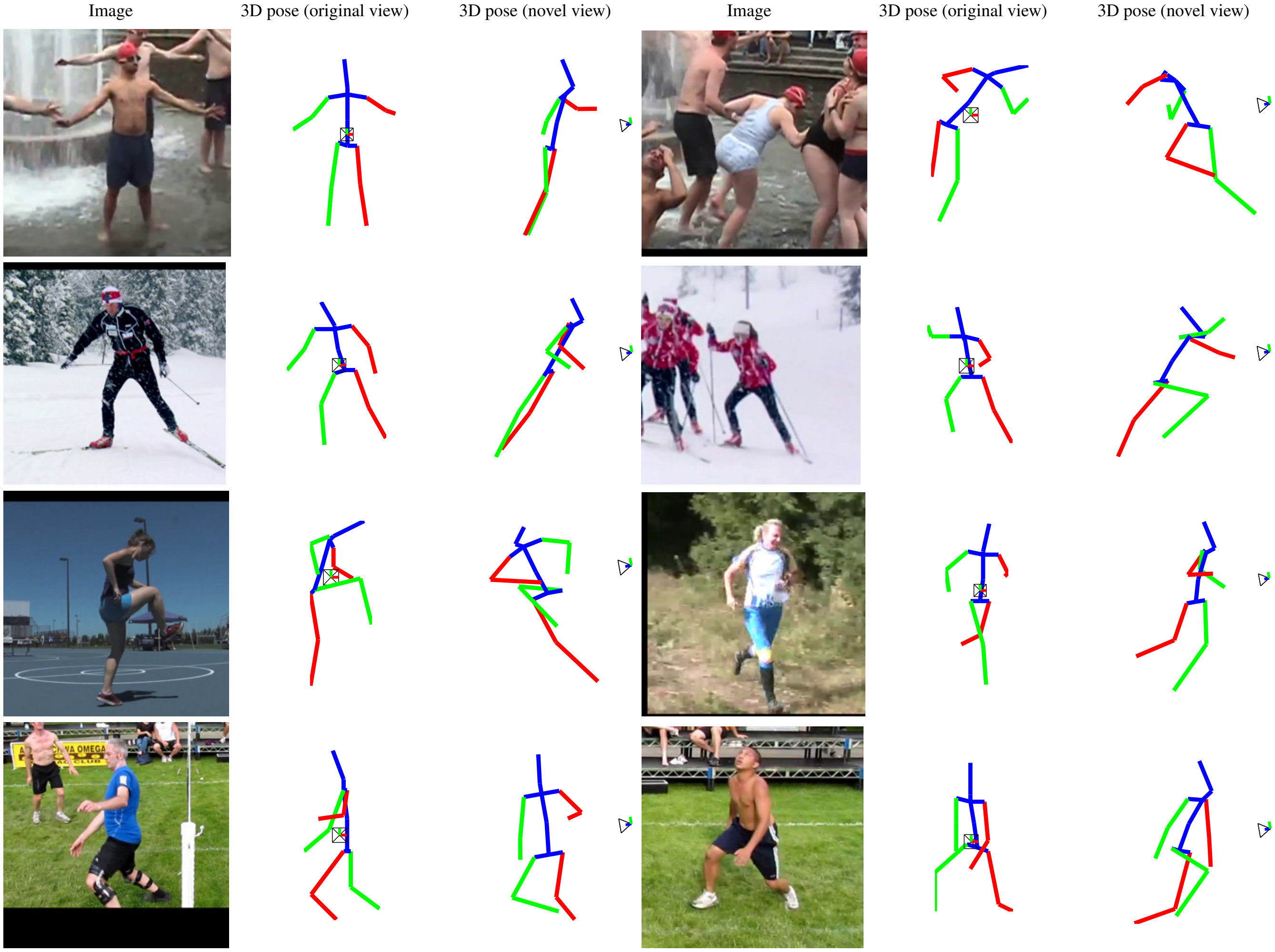

Figure 1: Successful reconstructions on MPII - Example set 1. For each example, we present (left-to-right) the input image, the predicted 3D pose from the original view, and a novel view. Notice the large variability in the image conditions and the articulated poses of the subjects.

[2] I. Kostrikov and J. Gall. Depth sweep regression forests for estimating 3D human pose from images. In *BMVC*, 2014. 1

[3] H. S. Park and Y. Sheikh. 3D reconstruction of a smooth articulated trajectory from a monocular image sequence. In *ICCV*, 2011. 1

[4] 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. 1

[5] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3D human pose estimation from monocular video. In *CVPR*, 2016. 2

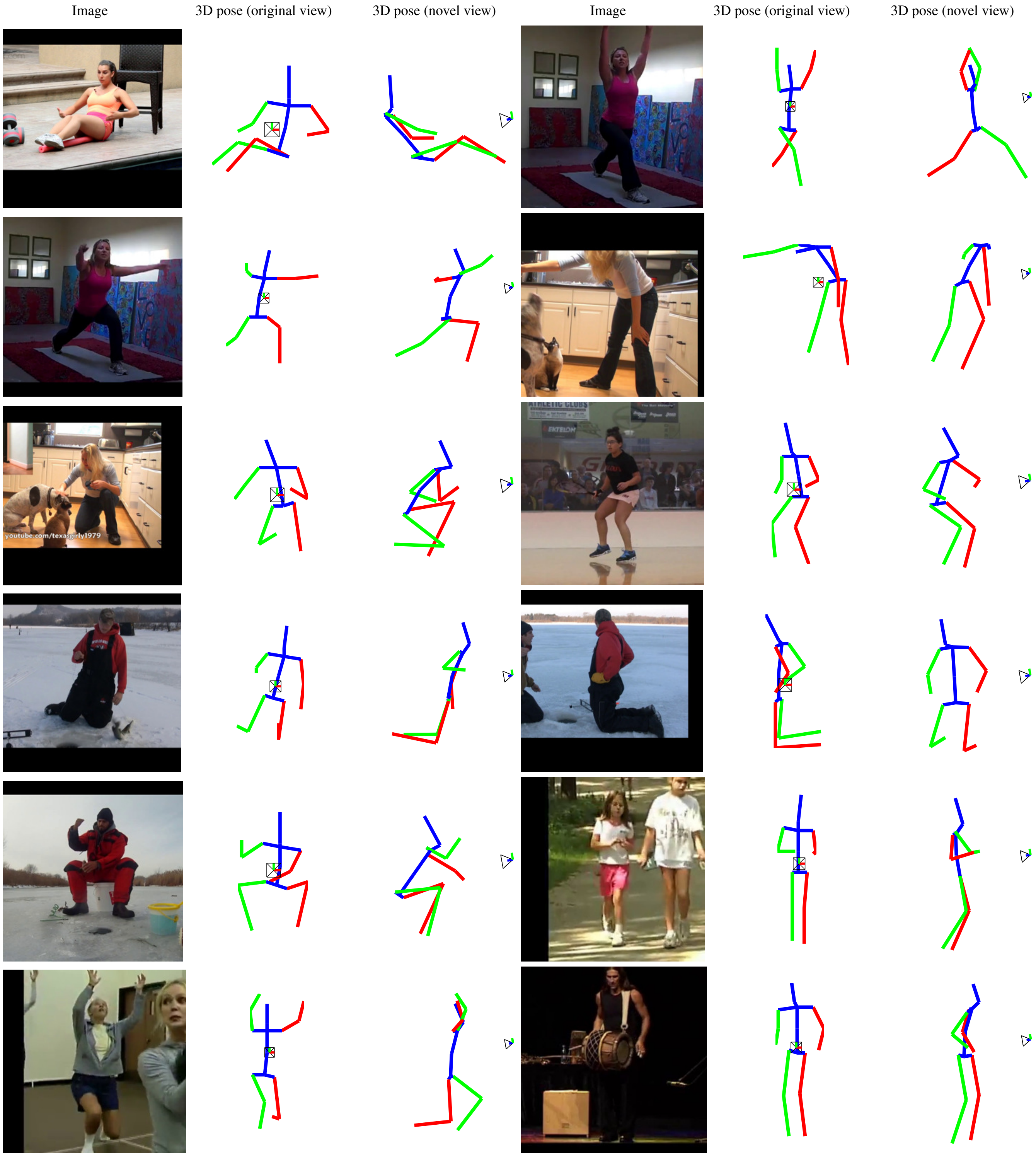

Figure 2: Successful reconstructions on MPII - Example set 2. For each example, we present (left-to-right) the input image, the predicted 3D pose from the original view, and a novel view. Notice the large variability in the image conditions and the articulated poses of the subjects.

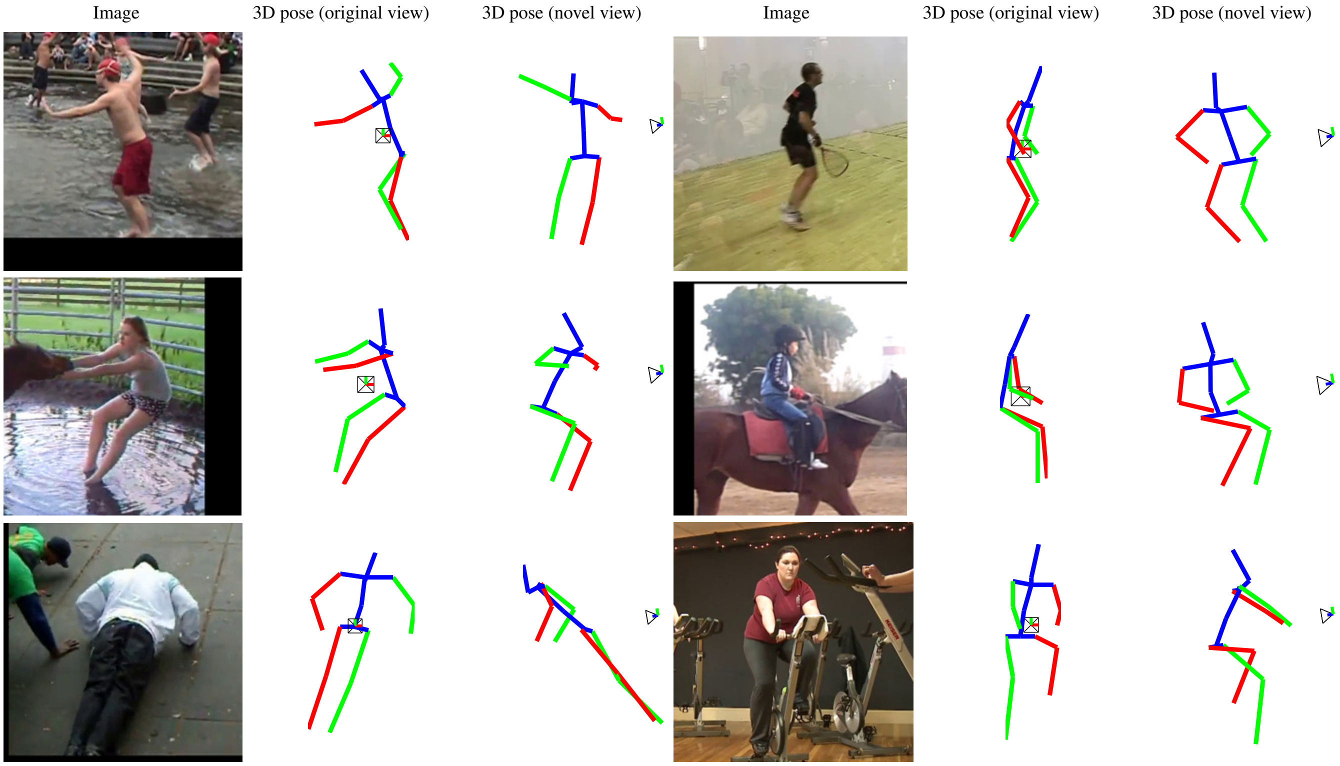

Figure 3: Successful reconstructions on MPII - Example set 3. For each example, we present (left-to-right) the input image, the predicted 3D pose from the original view, and a novel view. Notice the large variability in the image conditions and the articulated poses of the subjects.

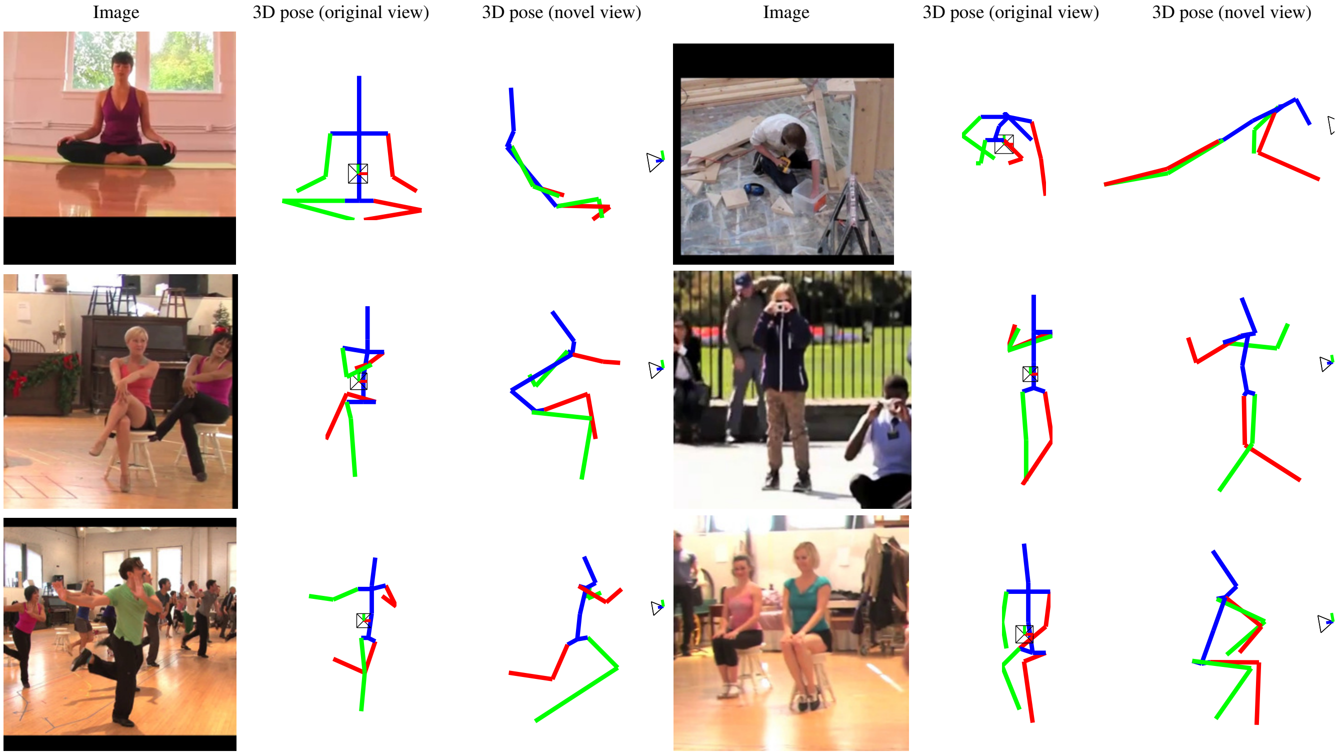

Figure 4: Erroneous reconstructions on MPII. For each example, we present (left-to-right) the input image, the predicted 3D pose from the original view, and a novel view. Failures can be attributed to challenging viewpoint, novel poses or severe self-occlusions.

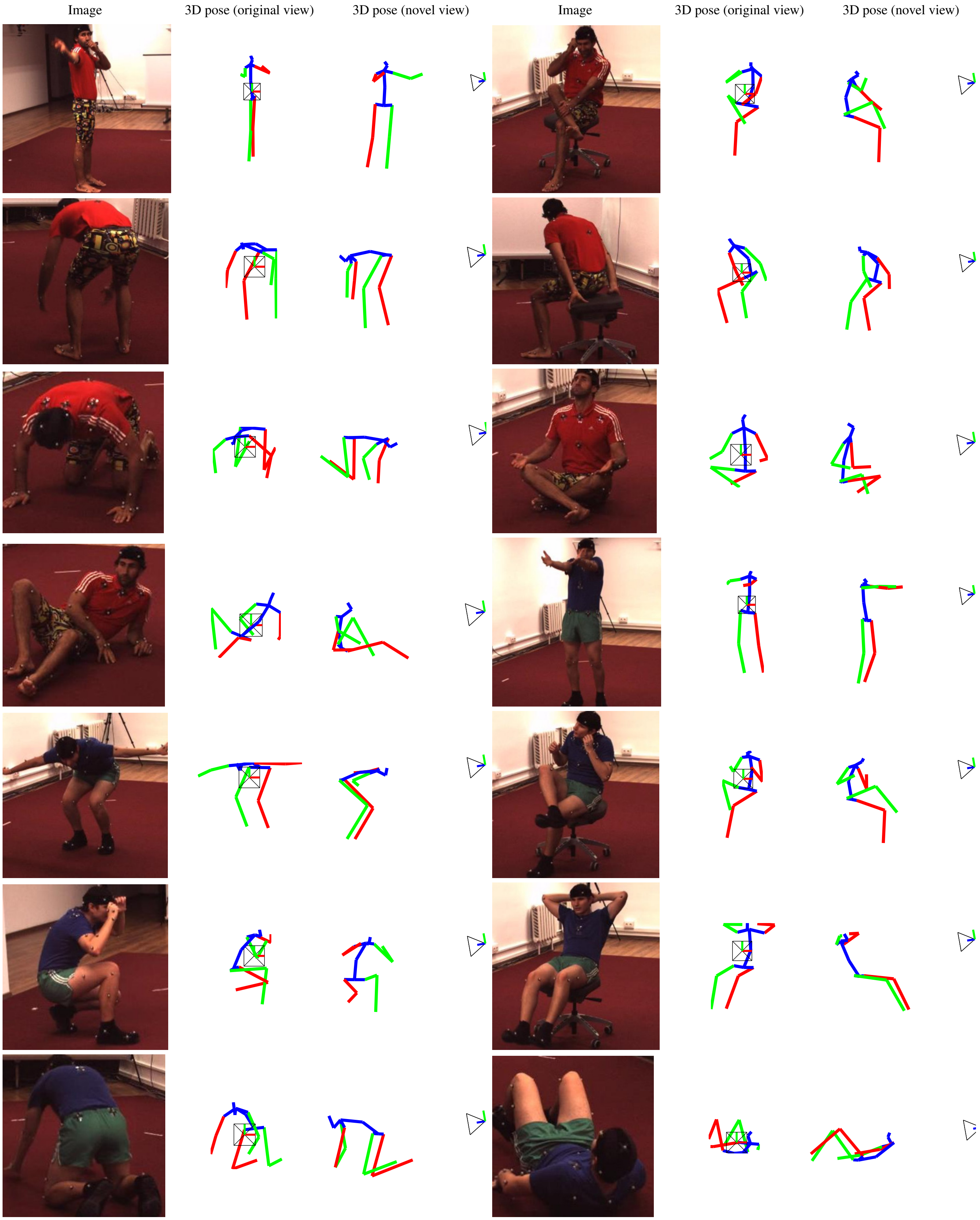

Figure 5: Qualitative results on Human3.6M. For each example, we present (left-to-right) the input image, the predicted 3D pose from the original view, and a novel view. Notice the large variability of the articulated poses.

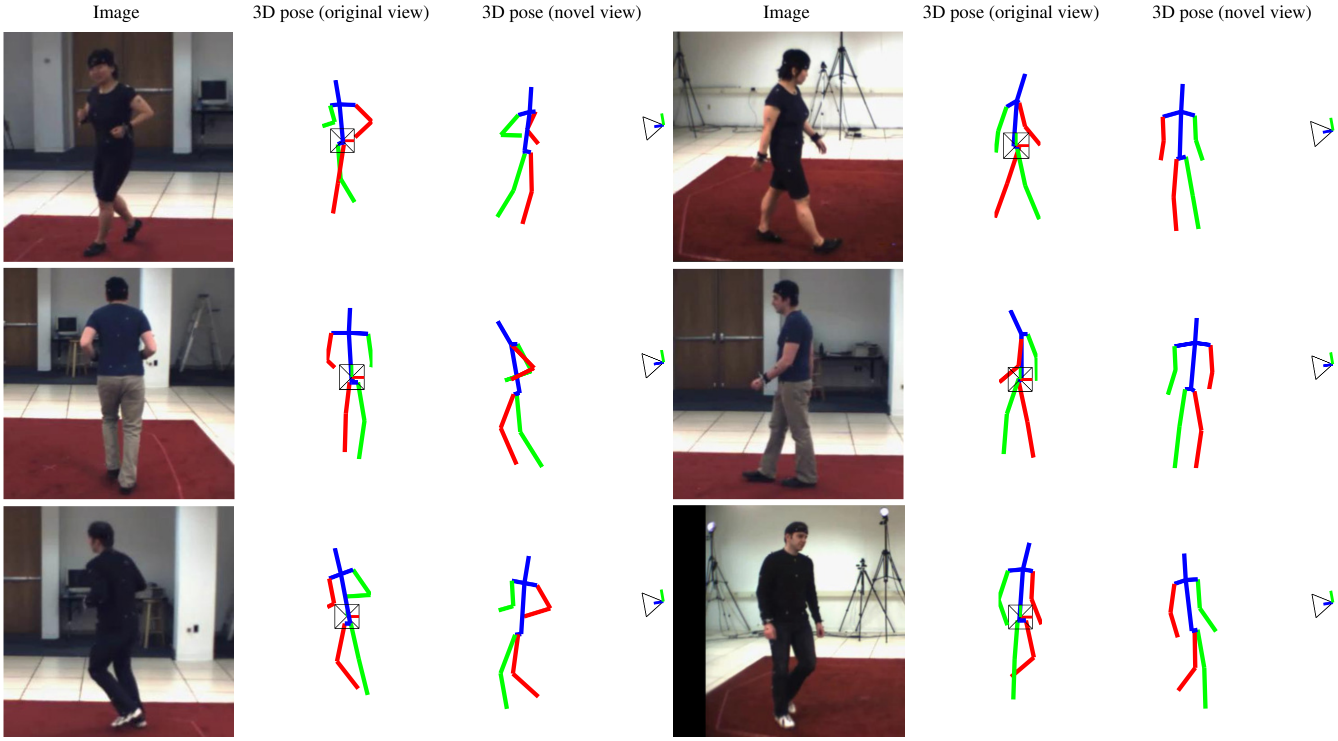

Figure 6: Qualitative results on HumanEva-I. For each example, we present (left-to-right) the input image, the predicted 3D pose from the original view, and a novel view.

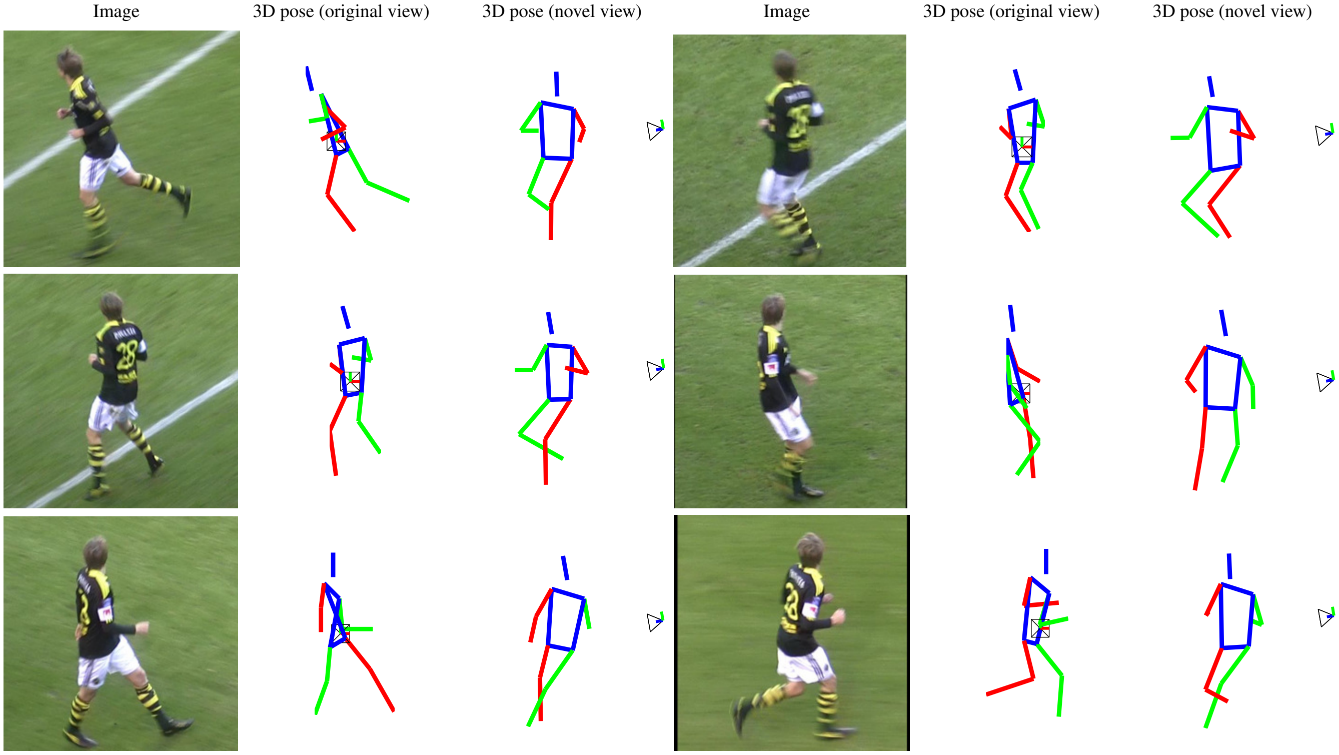

Figure 7: Qualitative results on KTH Football II. For each example, we present (left-to-right) the input image, the predicted 3D pose from the original view, and a novel view.

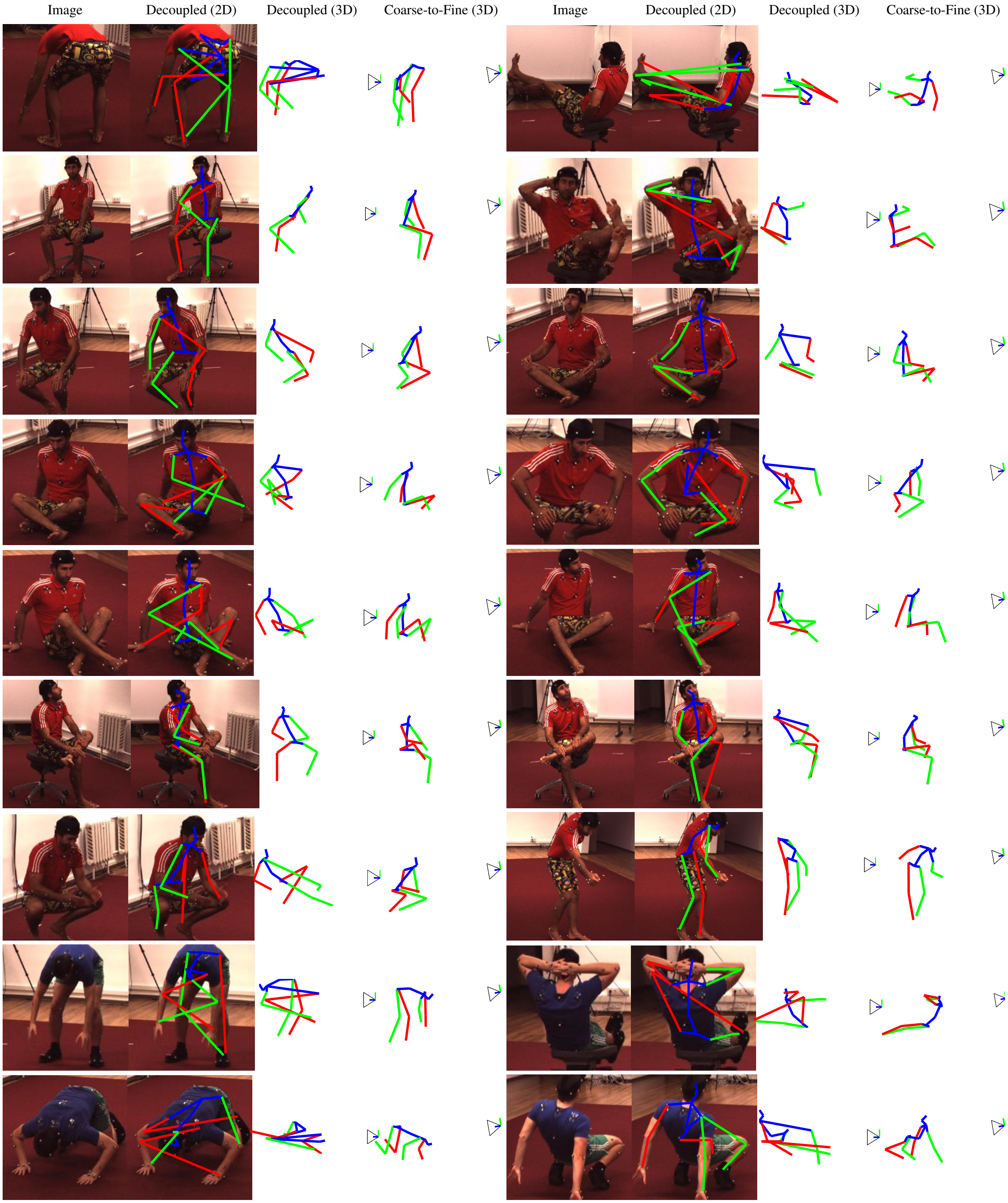

Figure 8: Results on Human3.6M using the "Decoupled" architecture versus our proposed "Coarse-to-Fine" approach (see Sections 3.3 and 4.4 of the main manuscript). For each example we present (left-to-right) the input image, the 2D and 3D result of the "Decoupled" architecture and the 3D result of our "Coarse-to-Fine". Failures for the "Decoupled" architecture can be attributed to erroneous or ambiguous initial 2D joint localization, which can lead the predicted 3D pose astray. The "Coarse-to-Fine" architecture instead uses image features end-to-end allowing the recovered 3D pose to rely on additional image evidence beyond 2D joint locations.