SkelSplat: Robust Multi-view 3D Human Pose Estimation with Differentiable Gaussian Rendering

Supplementary Materials

Laura Bragagnolo
University of Padova
bragagnolo@dei.unipd.it

Leonardo Barcellona
University of Amsterdam
l.barcellona@uva.nl

Stefano Ghidoni
University of Padova
ghidoni@dei.unipd.it

A. Pseudo Ground Truth Generation

This section provides additional details on how we generate the pseudo ground truth used during the optimization of *SkelSplat*. We begin by recalling that a human skeleton is defined as a set of joints $\text{SK} = \{sk_0, \dots, sk_N\}$, each of which is represented by an anisotropic 3D Gaussian. This yields a corresponding set of Gaussians $\text{GS} = \{gs_0, \dots, gs_N\}$, where each Gaussian gs_j encodes the spatial uncertainty around joint sk_j in 3D space. The optimization process leverages 2D keypoint detections obtained from a pre-trained 2D human pose estimator. Specifically, for each camera view i in a set of M synchronized and calibrated views (i.e., $i \in \{1, \dots, M\}$), we extract 2D keypoint locations $\text{SK}_\text{i}^\text{2D} = \{sk_{i0}^{2D}, \dots, sk_{iN}^{2D}\}$ corresponding to the projection of each joint in the image plane. To supervise the optimization with view-dependent supervision, we generate a set of pseudo ground truth heatmaps $\{I_{ij}\}_{j=1}^{N}$ for each camera view. Each heatmap $I_{ij} \in \mathbb{R}^{H \times W}$ represents a soft target for joint j in view i, and is constructed by rendering a 2D Gaussian gs_{ij}^{2D} centered at the detected 2D location $sk_{ij}^{2D} \in \mathbb{R}^2$. The shape of this Gaussian is determined by a covariance matrix $\Sigma_{ij}^{2D} \in \mathbb{R}^{2\times 2}$, which is obtained by projecting the original 3D covariance Σ_j of joint j into the 2D image plane of camera i as follows:

$$\Sigma_{ij}^{2D} = J_i W_i \Sigma_j W_i^T J_i^T \, , \tag{1}$$

where $W_i \in \mathbb{R}^{4\times 4}$ is the camera extrinsic transformation (world-to-camera), and $J_i \in \mathbb{R}^{2\times 3}$ is the Jacobian matrix of the perspective projection at the joint position in camera coordinates. In detail, if $\mu_j \in \mathbb{R}^3$ is the 3D joint position, its homogeneous coordinate $\tilde{\mu}_j = [\mu_j^\top, 1]^\top$ is transformed into camera coordinates as:

$$\tilde{\mu}_{ij}^{\text{cam}} = W_i \tilde{\mu}_j = \begin{bmatrix} X_{ij} \\ Y_{ij} \\ Z_{ij} \\ 1 \end{bmatrix} . \tag{2}$$

The Jacobian J_i for the perspective projection with focal lengths $f_{x,i}, f_{y,i}$ is:

$$J_i = \begin{bmatrix} \frac{f_{x,i}}{Z_{ij}} & 0 & -\frac{f_{x,i} X_{ij}}{Z_{ij}^2} \\ 0 & \frac{f_{y,i}}{Z_{ij}} & -\frac{f_{y,i} Y_{ij}}{Z_{ij}^2} \end{bmatrix} . \tag{3}$$

A small constant h is added to the covariance to prevent numerical issues. To characterize the shape of the 2D covariance ellipse, we compute the eigenvalues λ_1 and λ_2 of Σ_{ij}^{2D}, which correspond to the principal axes variances:

$$\det = \Sigma_{ij}^{2D}(1,1) \cdot \Sigma_{ij}^{2D}(2,2) - \left(\Sigma_{ij}^{2D}(1,2)\right)^2 , \tag{4}$$

$$m = \frac{\Sigma_{ij}^{2D}(1,1) + \Sigma_{ij}^{2D}(2,2)}{2} , \tag{5}$$

$$\lambda_1 = m + \sqrt{\max(\epsilon, m^2 - \det)} , \tag{6}$$

$$\lambda_2 = m - \sqrt{\max(\epsilon, m^2 - \det)} , \tag{7}$$

where ϵ is a small positive constant to ensure numerical stability, (s, q) indicates the element with s, q coordinates in the matrix.

B. Additional Details on Ablation Studies

In this section, we present further details on the ablation studies discussed in the main paper. Specifically, we include expanded analyses of noise robustness (as shown in Fig. 1), loss component contributions (see Tab. 2), covariance scaling behaviors (see Tab. 3), and the effects of rendering resolution (see Tab. 1).

Robustness to noisy initialization On Human3.6M [2], we perturb triangulation of MeTRAbs [3] and ResNet-152 poses with Gaussian noise of increasing standard deviation applied independently to each joint. We consider values for σ equal to 10, 20, 40, 60, 80 and 100 mm. From Fig. 1 we can observe how performance remains stable up to 40 mm noise and starts to degrade at 60 mm. With strong noise (80-100 mm) accuracy drops more sharply.

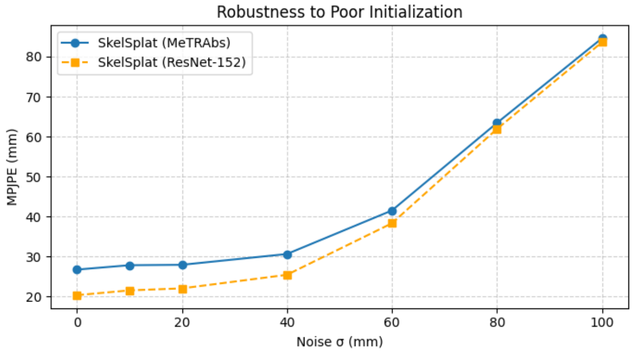

Figure 1. Ablation on robustness to poor initialization, adding Gaussian noise to triangulated joints.

Image resolution scale (%)	100	75	50	25
SkelSplat (MeTRAbs [3])	**26.7**	**26.7**	26.8	27.5
SkelSplat (ResNet-152)	**20.3**	20.4	20.7	21.7

Table 1. Resolution ablation. We downscale inputs to a percentage of the original size before rendering; higher resolution yields slightly lower MPJPE (mm).

Contribution of 3D loss to optimization Results in Tab. 2 show results for *SkelSplat* using three variants of 3D symmetry loss during optimization: Symm-1 applies the symmetry constraint to the lower arms and lower legs, Symm-2 extends the constraint to the upper arms and legs and Symm-3 further adds constraints from the hip joints to the root and from the shoulders to the neck. Tab. 2 illustrates accuracy on Human3.6M [2], Human3.6M-Occ-2 and Human3.6M-Occ-3 [1].

Effect of rendering resolution We assess the impact of rendering resolution on pose accuracy by reducing the input images to 75%, 50%, and 25% of their original resolution. Tab. 1 reports full results for *SkelSplat* using 2D detections from MeTRAbs and ResNet-152.

Impact of covariance scaling for 2D pseudo ground truth generation We evaluate the effect of enlarging the covariance for frequently occluded joints, such as elbows, hands, and knees, on Human3.6-Occ-2 and Human3.6-Occ-3. Tab. 3 reports results using scaling factors of 1.25, 1.5, and 2 applied to the default covariance of these joints. Overly large covariances decrease reconstruction accuracy, since the optimization becomes too tolerant of noisy or imprecise 2D inputs.

C. Joint Covariance for Confidence Estimation

While our experiments primarily evaluate the 3D means of the Gaussians as joint predictions, the associated covariances also encode potentially useful information about prediction uncertainty. To assess this, we compute the percentage of ground-truth joints that fall within 1, 2, or 3 standard deviations of the predicted Gaussian means on the Human3.6M dataset and on its occluded version Human3.6M-Occ-3. Here, in Fig. 2 and Fig. 3 we include results for joint-wise coverage in both occluded and non-occluded settings. Notably, in both cases, joints that are often occluded or self-occluded, such as hands, elbows, and ankles, tend to exhibit lower coverage.

Absolute MPJPE, mm					
Human3.6M	MeTRAbs [3]	27.0	**26.7**	26.9	26.8
	ResNet-152	20.6	**20.3**	20.4	20.4
Human3.6M-Occ-2	MeTRAbs [3]	30.0	29.6	29.5	**29.4**
	ResNet-152	26.2	24.6	23.8	**23.7**
Human3.6M-Occ-3	MeTRAbs [3]	31.4	31.1	**31.0**	**31.0**
	ResNet-152	27.2	27.0	**26.0**	26.1
3D Symm-1		-	✓	✓	✓
3D Symm-2		-	-	✓	✓
3D Symm-3		-	-	-	✓
Sec/iter		**0.028**	0.039	0.045	0.056

Table 2. Ablation study on different 3D symmetry loss contributions on the Human3.6M dataset and its occluded versions.

Scaling factor		1.25	1.5	2.0
H3.6M-Occ-2	MeTRAbs [3]	**29.6**	31.0	60.5
	ResNet-152	**24.6**	26.0	54.3
H3.6M-Occ-3	MeTRAbs [3]	**31.1**	32.6	61.9
	ResNet-152	**27.0**	28.9	59.0

Table 3. Ablation study on different methods to initialize 3D joint positions, absolute MPJPE (mm).

D. Supplementary Visualizations

In Fig. 4, we provide a visualization for a scene with four camera views from Human3.6M-Occ-3. We show the aggregated pseudo ground truth heatmap, obtained by summing the per-joint 2D heatmaps across all joints, together with a comparison between the initial and the final (optimized) 3D joint Gaussians. This highlights how the optimization step progressively refines the 3D pose, leading to improved alignment with the set of multi-view 2D detections and producing a more coherent reconstruction.

In Fig. 5, we report additional qualitative results on Human3.6M, Human3.6M-Occ-3, and the CMU Panoptic Studio datasets. For Human3.6M-Occ-3, we show only one of the occluded camera views. final row presents representative failure cases, in which our method struggles due to factors such as extreme occlusion or complex limb configurations, for which inconsistent multi-view evidence is common.

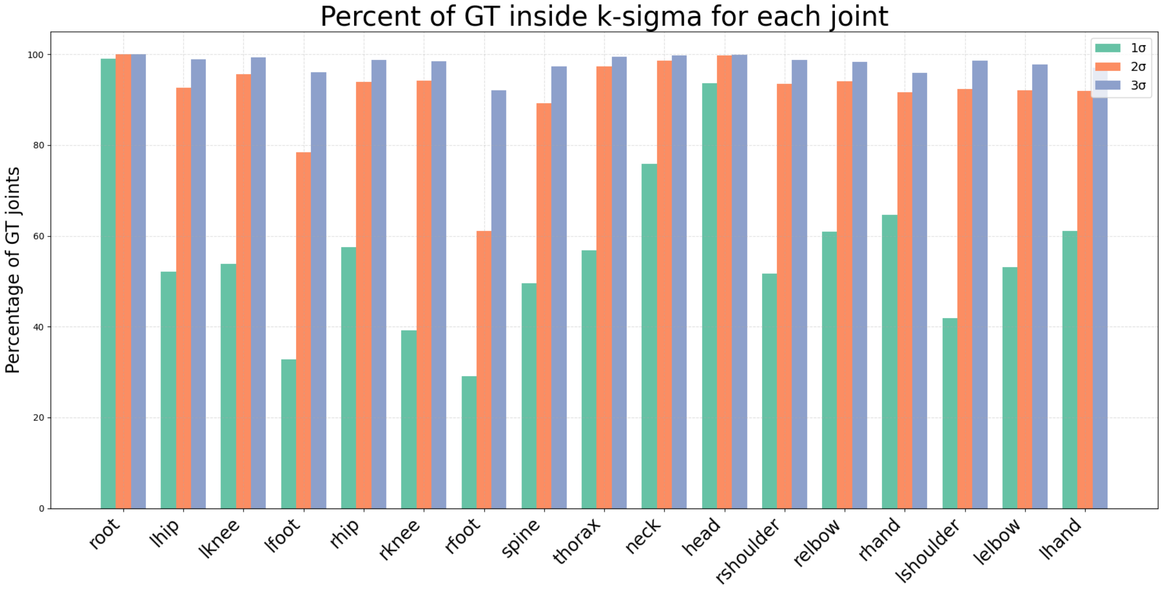

Figure 2. Joint-wise coverage rates across different sigma thresholds for the Human3.6M dataset.

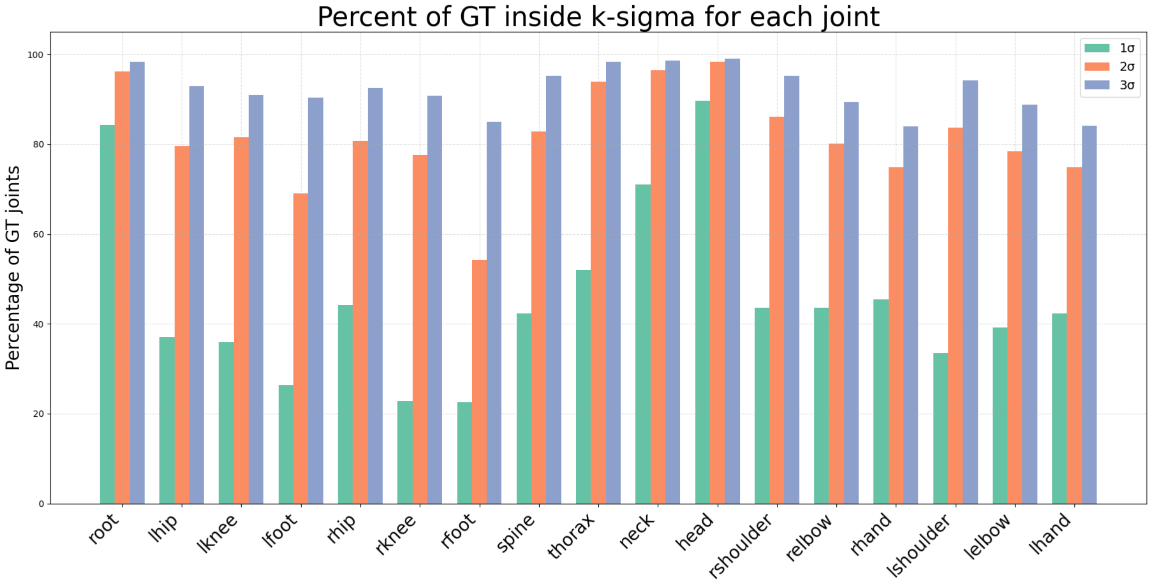

Figure 3. Joint-wise coverage rates across different sigma thresholds for the Human3.6M-Occ-3 dataset.

References

[1] Laura Bragagnolo, Matteo Terreran, Davide Allegro, and Stefano Ghidoni. Multi-view pose fusion for occlusion-aware 3d human pose estimation. In *European Conference on Computer Vision*, pages 117–133. Springer, 2025. 2

[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. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1, 2

[3] István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3d human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2956–2966, 2023. 1, 2

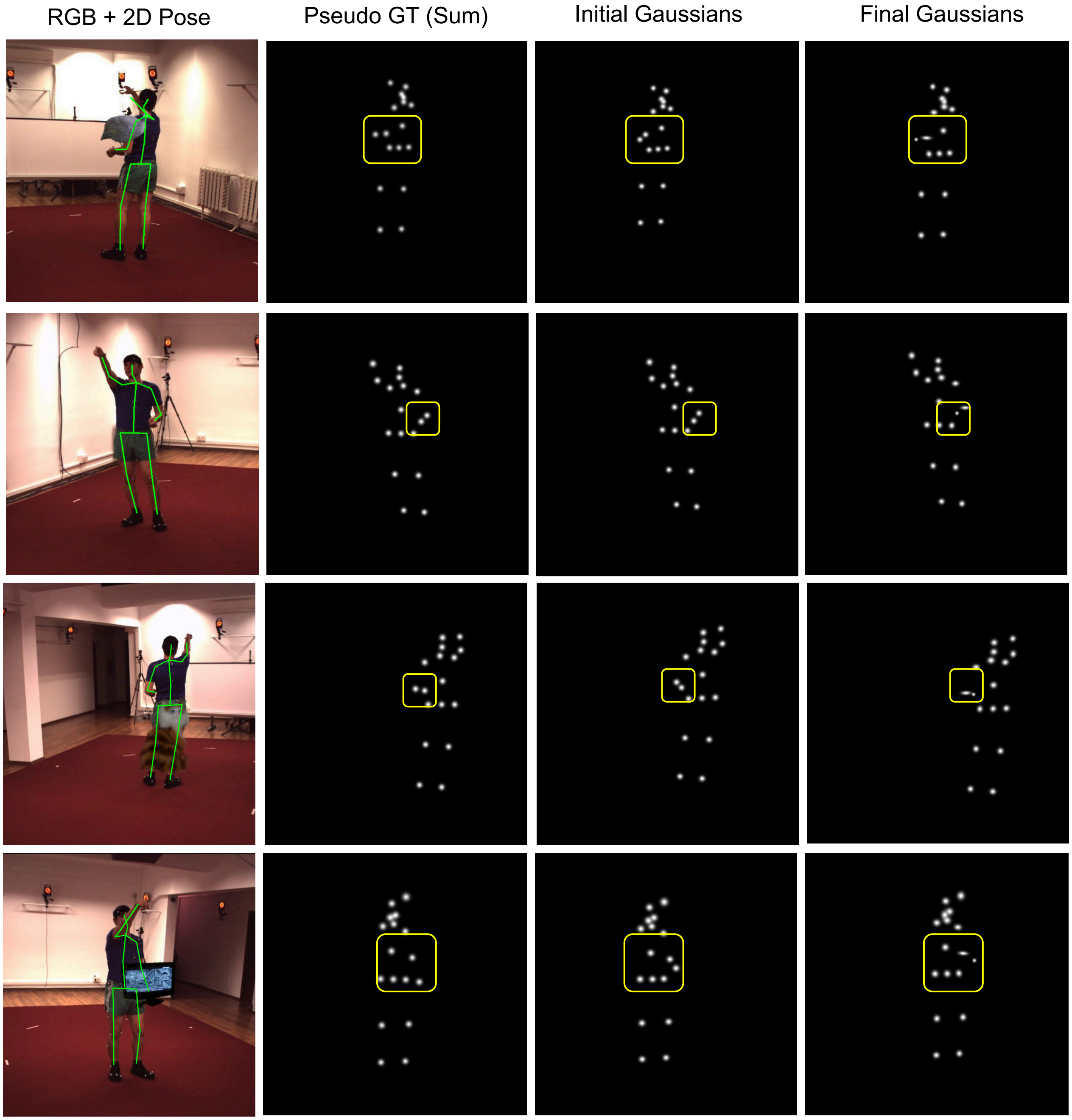

Figure 4. Visualization of pseudo ground-truth supervision and joint refinement. For 4 camera views, we show (left) the aggregated pseudo ground-truth heatmap obtained by summing the 2D Gaussian heatmaps of all joints, and (right) a 3D visualization of the joint Gaussians before and after optimization.

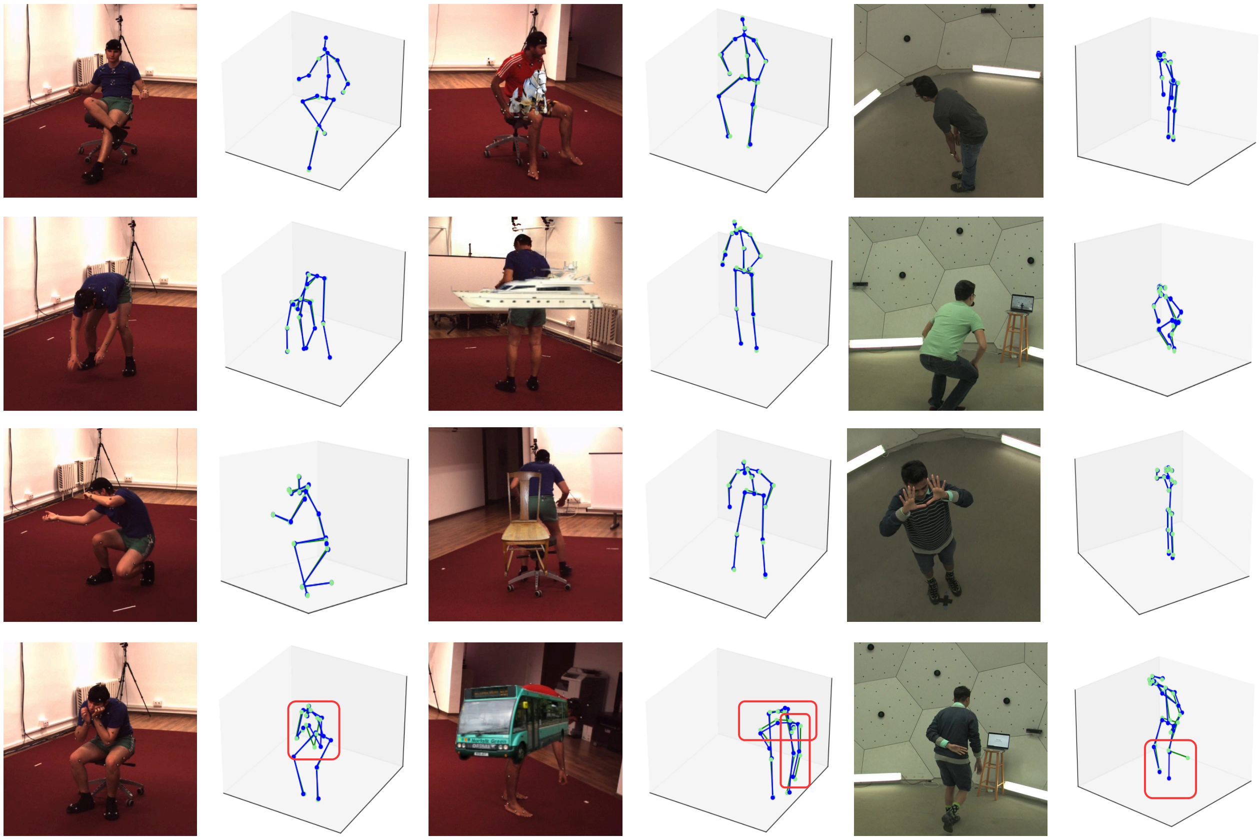

Figure 5. Qualitative results on Human3.6M (left), Human3.6M-Occ-3 (middle), and CMU Panoptic (right). For Human3.6M-Occ-3 we show one of the three occluded views. The last row shows some failure cases.