Robust Crowd Reconstruction

A Preprint

Briefly

The paper addresses the challenging problem of reconstructing 3D human poses, shapes, and global positions for hundreds of people from a single large-scene image captured under arbitrary camera fields of view. Unlike conventional multi-person reconstruction methods that assume relatively large human scales or fixed camera parameters, this work explicitly targets surveillance-like scenarios where humans appear small, highly variable in scale, and subject to severe perspective distortion.

To tackle these challenges, the authors propose Robust Crowd Reconstruction (RCR), a top-down pipeline that combines detection-based processing with a novel geometric formulation. A central idea is the introduction of the Human-scene Virtual Interaction Point (HVIP), which converts the ambiguous 3D localization problem into a 2D estimation task constrained by the ground plane. This formulation allows the method to recover globally consistent camera-space reconstructions without relying on test-time optimization.

The pipeline further incorporates an Iterative Ground-aware Cropping strategy to reliably detect humans across extreme scale variations, and introduces a canonical Upright 2D/3D Space to decouple reconstruction from camera parameters and cropping artifacts. Together, these components enable accurate reprojection, plausible human-ground interactions, and stable generalization across different camera FoVs.

Human-scene Virtual Interaction Point (HVIP)

The HVIP concept is a key contribution of the paper and serves as the foundation for resolving depth ambiguity in monocular large-scene reconstruction. Instead of directly regressing absolute 3D positions, each person is associated with a virtual interaction point defined as the projection of the torso center onto the ground plane. By estimating this point in the image plane, the method effectively reduces 3D localization to a constrained 2D prediction problem.

This design choice is particularly well-suited to large scenes, where absolute depth is difficult to infer but relative alignment with the ground is often reliable. Given estimated camera intrinsics and ground parameters, the HVIP allows the recovery of consistent 3D torso centers while maintaining correct reprojection. The formulation is geometrically grounded and avoids heuristic depth scaling commonly used in prior work.

Iterative Ground-aware Cropping

A major practical challenge in large-scene images is detecting humans whose pixel sizes vary drastically across the image. The authors address this with an iterative cropping mechanism guided by estimated ground geometry. Rather than relying on fixed or manually tuned crop sizes, the method alternates between detection, camera-and-ground estimation, and adaptive cropping.

This iterative process resolves the circular dependency between detection quality and geometric estimation. Initial coarse crops enable rough keypoint detection, which in turn supports more accurate ground estimation. The refined ground model then guides subsequent cropping, ensuring that each person appears at a suitable scale in at least one crop. This approach significantly improves recall while avoiding duplicate detections, and removes the need for scene-specific hyperparameter tuning.

Upright 2D and 3D Spaces

To eliminate the influence of camera parameters and perspective distortion during single-person reconstruction, the paper introduces a canonical Upright 2D Space and a corresponding Upright 3D Space. In these spaces, the ground normal is aligned with the vertical axis and an orthographic projection model becomes applicable.

Each detected human crop is transformed into the Upright 2D Space via a learned homography derived from camera and ground estimates. Single-person pose and shape estimation is then performed in the Upright 3D Space using standard weak-perspective SMPL regressors. This normalization greatly simplifies learning, as the reconstruction network no longer needs to account for arbitrary camera FoVs or local perspective effects.

The use of Upright Normalization represents a clean separation between geometry-driven alignment and learning-based reconstruction, and is shown to be critical for both pose accuracy and reprojection consistency.

HVIPNet and Reconstruction

The paper introduces HVIPNet, a lightweight network dedicated to predicting the relative vertical offset between the torso center and the HVIP in the Upright 2D Space. Since scale and orientation are normalized, this estimation becomes significantly more stable than predicting depth directly in camera space.

After estimating SMPL parameters and HVIPs in the upright representation, all reconstructed humans are transformed back into a unified camera space. This final step yields globally consistent 3D poses and shapes, correct relative positioning among individuals, and accurate reprojection into the original image.

Datasets and Evaluation

To support training and evaluation, the authors contribute two datasets. LargeCrowd provides real-world large-scene images with extensive 2D annotations, while SynCrowd offers synthetic data with full 3D ground truth under varying camera FoVs. Together, these datasets enable both qualitative and quantitative evaluation of spatial consistency, reprojection accuracy, and pose quality.

Experimental results demonstrate that RCR substantially outperforms prior methods in terms of spatial arrangement, depth ordering, and robustness to camera variation. Ablation studies further confirm the importance of Upright Normalization and iterative ground-aware cropping.

Summary

Overall, the paper presents a carefully designed system that combines geometric insight with practical engineering choices. By reformulating depth estimation through HVIP, normalizing reconstruction via upright spaces, and addressing scale variation through iterative cropping, RCR achieves reliable crowd reconstruction in scenarios that are challenging for existing methods. The approach is conceptually clear, modular, and well-aligned with the constraints of real-world surveillance imagery.