The One Where They Reconstructed 3D Humans and Environments in TV Shows

Georgios Pavlakos * , Ethan Weber * , Matthew Tancik, Angjoo Kanazawa
University of California, Berkeley

Presented by
Xinyu Liu, Bang Gong

Reconstructing 3D Humans and Environments in TV Shows

Motivation

Goal:

For machines to aggregate 3D information over videos and perceive the 3D human pose and locations like human brains do

Why TV shows?

- repetitive environment throughout the shows
- repetitive characters

Related Works

TV Show Modeling

- 2D, without camera calibration

Scene Reconstruction

- most NeRF-based methods are limited when handling changes in appearances and transient objects
- NeRF-W can handle the problems

3D Human-Scene Interaction

- need environment reconstructed as *a priori*
- focus on visible human-object interactions

Overview

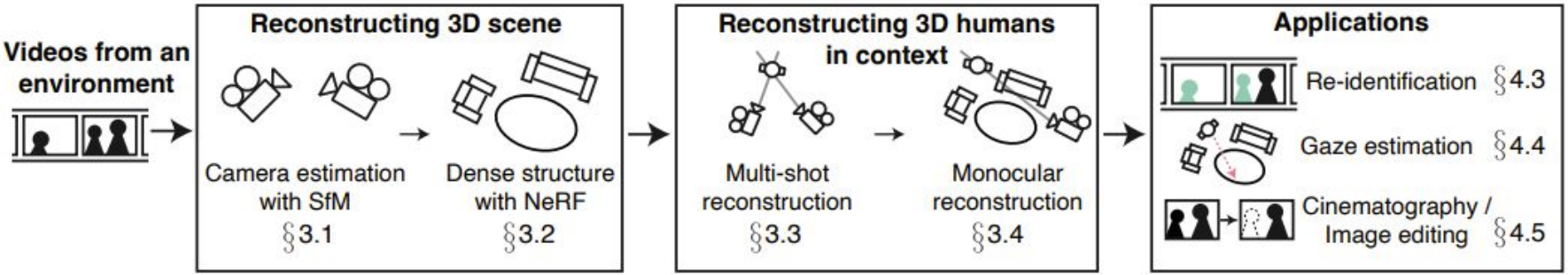


Fig. 2. Overview of our workflow. First, we use a collection of videos from a TV show environment and reconstruct the 3D scene (cameras and dense structure). We then use this information to recover accurate 3D pose and location of people over shot boundaries and on monocular frames. The recovered 3D information is immediately useful for various downstream applications.

Camera Registration

Frames sampled at shot boundaries

INPUT

- DISK (DIScrete Keywords) to find correspondences

- Mask R-CNN to detect human masks

- COLMAP to match remaining features, estimate sparse 3D reconstruction, and register cameras

Estimates of camera intrinsics $K_t \in R^{3\times3}$ and extrinsics $R_t^{CW} \in R^{3\times3}$, $T_t^{CW} \in R^3$, where CW denotes camera to world transformation.

OUTPUT

Dense Structure

- Traditional method: assume static scenes
- NeRF-W network for dense structure estimation

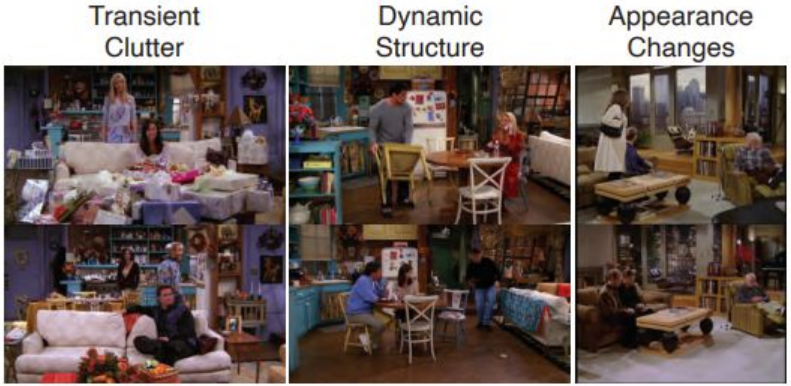
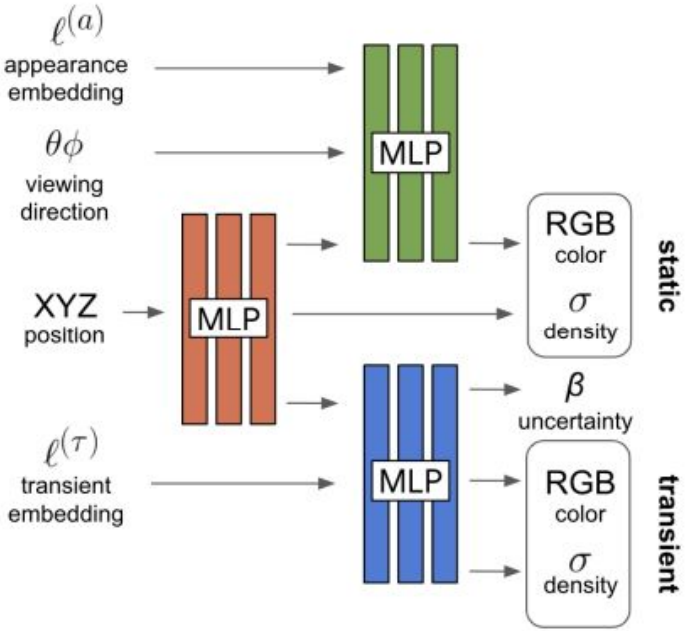


Fig. 3. Reconstruction challenges of TV shows include transient and dynamic objects as well as appearance changes.



Dense Structure

- Training Data: Select the image with least percent of Mask R-CNN human pixels in each cluster based on camera location and viewing direction

Fig. 4. Panoramic views of the reconstructed TV show environments. We obtain and render the static structure using NeRF-W [36]. The environments represent seven TV shows: "The Big Bang Theory", "Frasier", "Everybody Loves Raymond", "Friends", "Two And A Half Men", "Seinfeld" and "How I Met Your Mother".

Calibrated Multi-Shot Human Reconstruction

- Effective multi-view information requires knowledge of the identity of the actors across the shot changes
- Prior: a pre-trained recognition-based re-ID model
- Our calibrated multi-shot optimization: solve jointly 3D human pose, shape, location, and camera information

Fig. 5. Calibrated cameras for scale estimation and identity association. Given calibrated cameras, we can use frames at a shot change to solve for the actors' pose, location, relative scale and association. The four overlapping regions (left) indicate possible locations triangulated by the cameras. Circles indicate correct matches after Hungarian matching. Reconstructed humans are visualized in a NeRF (right).

SMPLify Fitting

- Minimize the objective function with respect to $\{\theta_t, \theta_{t+1}, \beta_t, \beta_{t+1}, T_t^C, T_{t+1}^C\}$:

$$E = \underbrace{E_{J_t} + E_{J_{t+1}}}_{\text{2D reprojection}} + \underbrace{E_{\text{priors}_t} + E_{\text{priors}_{t+1}}}_{\text{anthropometric constraints}} + \underbrace{E_{\text{glob}_{t,t+1}}}_{\text{3D consistency}} \tag{1}$$

$$E_{\text{glob}_{t,t+1}} = \|(R_t^{CW} J_t^C + T_t^{CW}) - (R_{t+1}^{CW} J_{t+1}^C + T_{t+1}^{CW})\|^2. \tag{2}$$

- Solve the association by Hungarian matching

Contextual Monocular Human Reconstruction

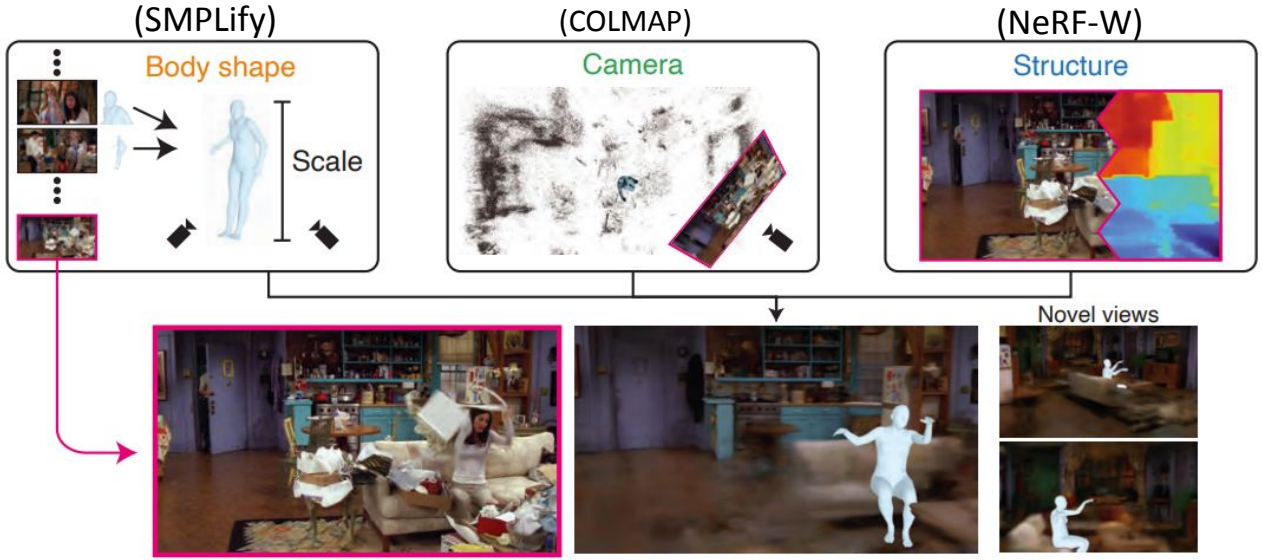


Fig. 6. Contextual monocular human reconstruction. For an input frame, we can leverage (a) the **body shape** (scale) of the person from a neighboring shot change, (b) the **camera** registration, and (c) the static **structure** of the environment. This enables monocular reconstruction of the person in context with their environment.

SMPLify Optimization

- we formulate an objective to discourage the human body vertices V from occupying areas with high density values:

$$E_{\mathbf{structure}} = \rho\Big(\sum_{v \in V} \tilde{\sigma}(v)\Big), \tag{3}$$

where ˜σ samples values from the density field σ using trilinear interpolation, while ρ is the Geman-McClure robust error function

Eventually, our monocular fitting objective minimizes:

$$E_J(\beta = \hat{\beta}, \theta, K = \hat{K}, J_{est}) + E_{\text{priors}} + E_{\mathbf{structure}}, \tag{4}$$

E_J with respect to shape parameters $\hat{\beta}$, pose parameters θ, camera intrinsics $\hat{K}$, 2D keypoints J_{est}

Applications

- Re-identification
- Gaze information
- Cinematography applications

Experiments

Training Set:

7 shows, 1 season per show, each environment has 1k-5k frames from shot change

Test Set:

50 person identities selected from each show on shot changes

Experiments:

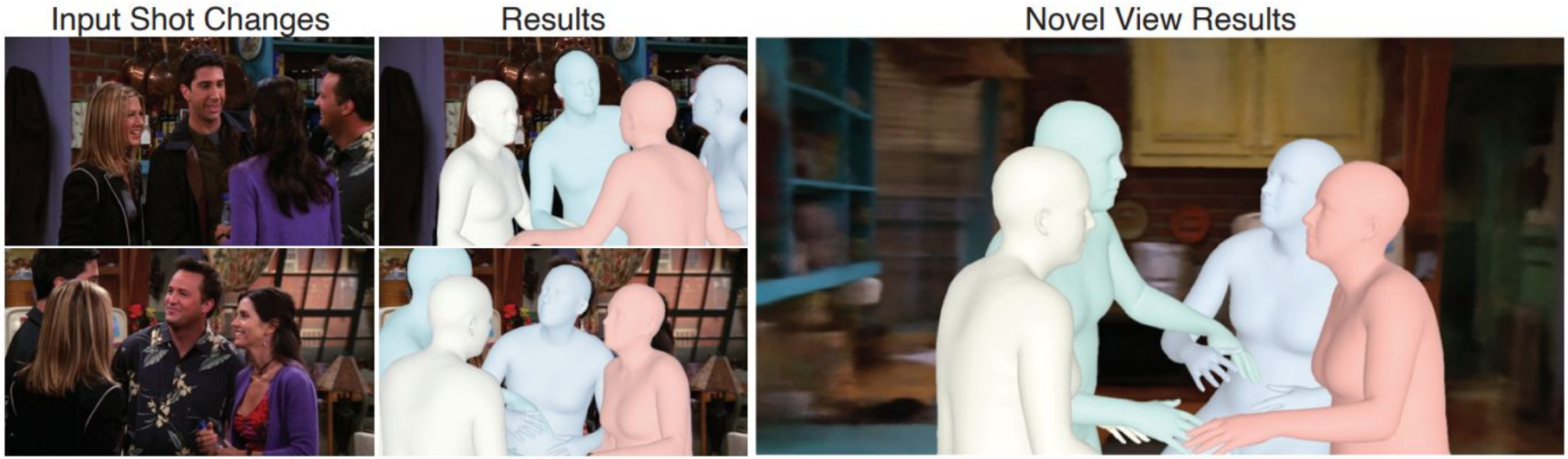
calibrated multi-shot human reconstruction, monocular contextual human reconstruction, re-identification, gaze estimation, cinematography/image editing applications

1. Calibrated multi-shot human reconstruction

Method	Camera information		Human3.6M		TV shows	
Multi-shot optimization	Intrinsics	Extrinsics	MPJPE	PA-MPJPE	% preferred vs. Ours ↑	Distance error ↓
Uncalibrated [45]	✗	✗	131.9	56.9	4%	889cm
Partial Calibration	✓	✗	123.8	56.3	35%	59cm
Calibrated	✓	✓	**65.8**	**47.1**	—	**38cm**

Table 1. Evaluation of the proposed calibrated multi-shot optimization. We ablate the effect of camera information in multi-shot optimization. On Human3.6M, we report results on the standard 3D pose metrics in mm [70]. On our TV show data, we perform a system evaluation on AMT and provide quantitative results based on the spatial localization of the reconstructed person in the scene.

1. Calibrated multi-shot human reconstruction



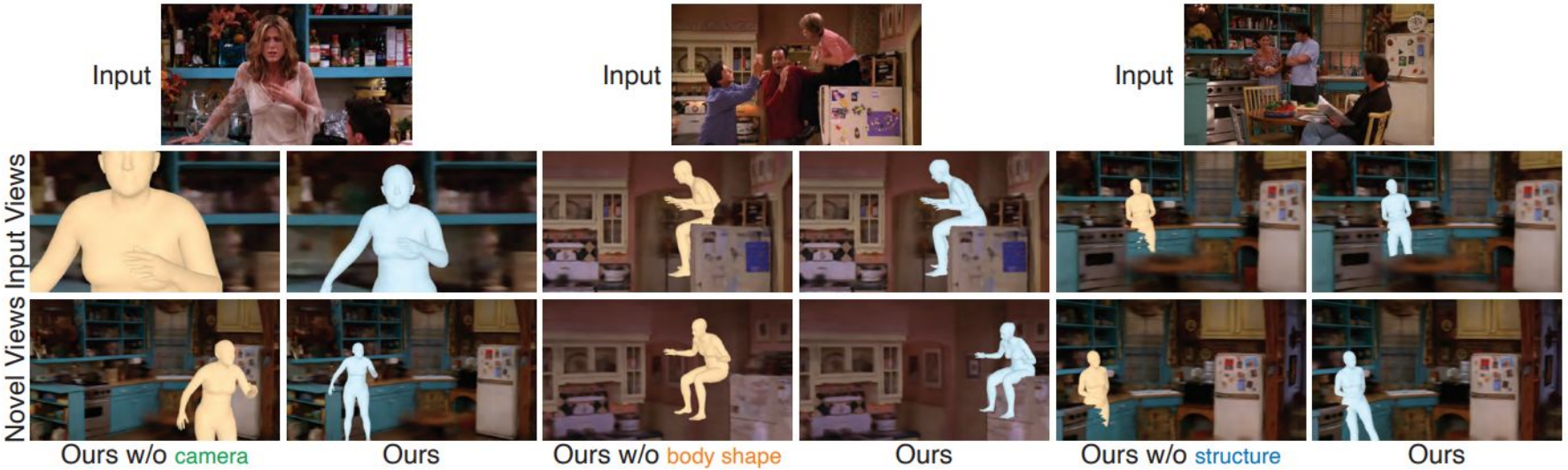
2. Monocular contextual human reconstruction

PCK = percentage of correct keypoints (between shots)

Method	cross-shot PCK
No context: ProHMR [28]	14.7%
No context: PARE [26]	14.2%
No context: SMPLify [5]	16.5%
Context w/o **camera** (intrinsics)	16.0%
Context w/o **body shape** (scale)	65.9%
Context w/o **structure**	87.5%
Context (full)	**88.7%**

Table 2. Ablation of the main components of our contextual reconstruction. Cross-shot PCK @ $\alpha = 0.5$ is reported. Knowledge of the camera focal length is very important to get a good 3D location for the human. Information about body shape can have significant improvements, as it resolves the scale ambiguity. Structure helps to avoid the incoherent interpenetrations with the scene.

2. Monocular contextual human reconstruction



3. Re-identification

Matching costs	Re-ID F1 ↑
Fu *et al.* [12] (Appearance)	0.78
Huang *et al.* [22] (Appearance)	0.79
Huang *et al.* [23] (Appearance)	0.80
Keypoint triangulation (Geometry)	0.86
Ours (Geometry + Anthropometric)	**0.91**

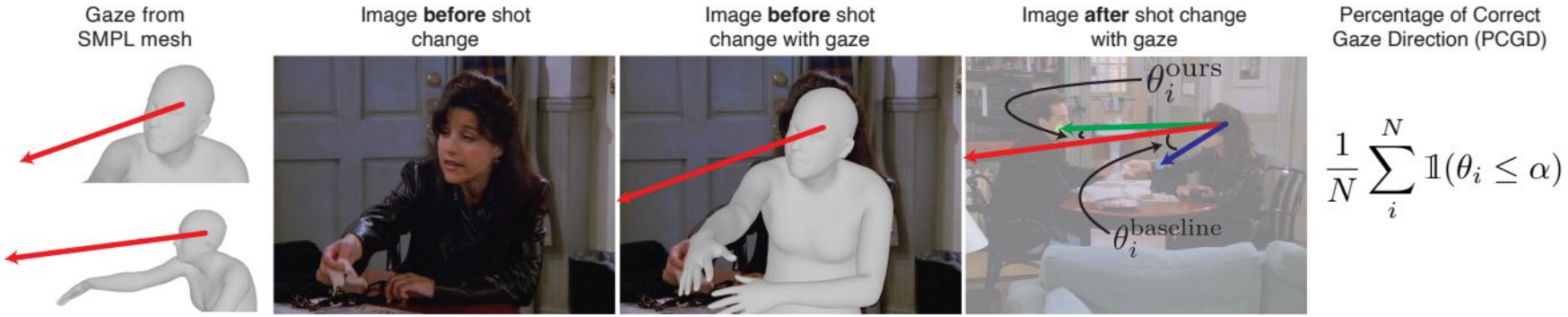
Table 3. Re-ID results for actors in shot boundary frames. We use different methods to estimate matching costs for detections and we run Hungarian matching to establish associations. A geometric baseline using the reprojection error from person keypoint triangulation improves upon SOTA image-based baselines [12,22,23], but using our multi-shot fitting cost performs better because it also includes anthropometric constraints, *i.e.*, the triangulated points should respect the human body priors.

Appearance: image-based networks for affinity estimation

Geometry: human keypoint triangulation based, uses recovered camera info

Anthropometric: uses human body shape as *a priori*

4. Gaze estimation



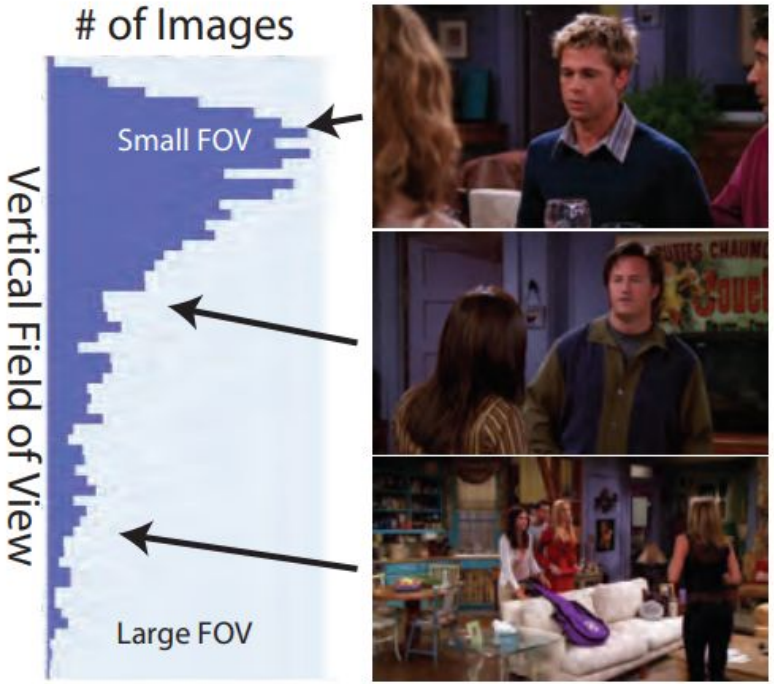
Percentage of Correct Gaze Direction (PCGD)

$$\frac{1}{N}\sum_{i}^{N}\mathbb{1}(\theta_i \leq \alpha)$$

Method	PCGD ($\alpha = 20^o$) ↑ all	w/ face
Recasens *et al.* [46]	16%	32%
Ours	**62%**	**67%**

Table 4. Gaze following results. We report the Percentage of Correct Gaze Directions (see text for description). Our approach outperforms the baseline of [46].

5. Cinematography/Image editing applications



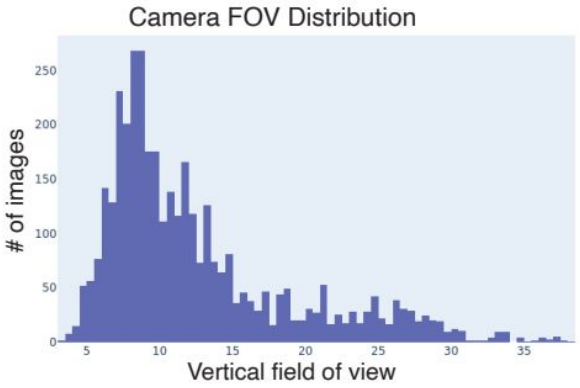
(a) Camera FOV Distribution

Close-ups provide small field of view while full-shots provide large field of view

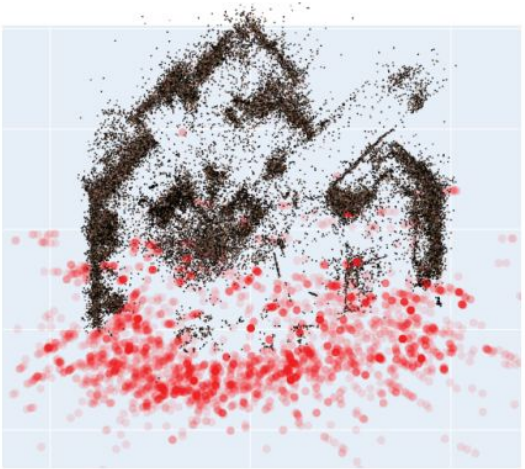
=> Justifies using data across whole season

5. Cinematography/Image editing applications

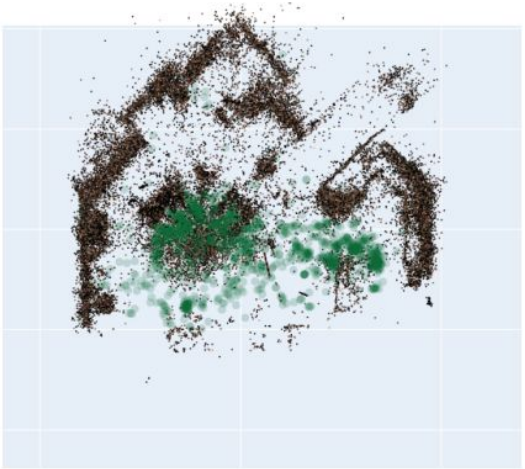
The Big Bang Theory: Sheldon's apartment



Camera Pose Distribution



Person Location Distribution



5. Cinematography/Image editing applications

(d) Person Removal

(e) The Big Bunny Insertion