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

Georgios Pavlakos*, Ethan Weber*, Matthew Tancik, Angjoo Kanazawa

University of California, Berkeley

Fig. 1. Reconstruction of humans in TV show environments. Given images across the whole season of a TV show, we present an approach that recovers the 3D scene context, which enables accurate estimation of every actor's 3D pose and location. We show the input (left), the mesh reconstructions of the actors in the camera view (center) and in a novel view (right). Human meshes are visualized against the reconstructed scene, which is represented by a Neural Radiance Field (NeRF). To appreciate the correct 3D localization of people, notice the position in the novel view and the occlusions. Readers are encouraged to watch video results in the project page: `http://ethanweber.me/sitcoms3D/`.

Abstract. TV shows depict a wide variety of human behaviors and have been studied extensively for their potential to be a rich source of data for many applications. However, the majority of the existing work focuses on 2D recognition tasks. In this paper, we make the observation that there is a certain persistence in TV shows, *i.e.*, repetition of the environments and the humans, which makes possible the 3D reconstruction of this content. Building on this insight, we propose an automatic approach that operates on an entire season of a TV show and aggregates information in 3D; we build a 3D model of the environment, compute camera information, static 3D scene structure and body scale information. Then, we demonstrate how this information acts as *rich 3D context*

* Equal contribution.

that can guide and improve the recovery of 3D human pose and position in these environments. Moreover, we show that reasoning about humans and their environment in 3D enables a broad range of downstream applications: re-identification, gaze estimation, cinematography and image editing. We apply our approach on environments from seven iconic TV shows and perform an extensive evaluation of the proposed system.

1 Introduction

Remember that time when you binge-watched an entire season of your favorite TV show, *e.g.*, "Friends", over a weekend? After that experience, you would know the layout of the rooms, the locations of the furniture, and even the relative height of the characters as they interact closely on screen. As a result, for any frame, you could tell where the room is viewed from, where the characters are situated, and how they relate to the rest of the scene, even the parts of the scene outside the frame. Essentially, as viewers, we aggregate all the visual information into a dynamic 3D world where the new observations are aligned to.

In this paper, we propose a method that can similarly aggregate 3D information over video collections and use it to perceive accurate 3D human pose and location of the actors. Although reconstruction of dynamic scenes is challenging from a single video clip, our insight is that in the context of TV shows, across many episodes, there are many video clips that *depict the same scene and people many times.* The repeated observations provide a strong multi-view signal of the underlying scene, enabling reconstruction of the camera and the dense structure. These serve as context to accurately recover the 3D pose and location of the people in the 3D environment. A representative result is shown in Figure 1. Although we demonstrate our method & results on TV shows, our insight is also applicable to other domains with repetition in the environment and the people, *e.g.*, sports [20,47,71], late night shows [15,40] and movies [45].

We operationalize our insight by focusing on an entire season from TV shows and collecting the sequences that correspond to a specific environment. These sequences are organized in shots [3], which are typically captured by different cameras. To collect a diverse set of images, we sample frames at the shot boundaries (cuts between cameras). This ensures a wide variety of viewpoints, while avoiding redundancy, making it practical to apply a structure-from-motion pipeline [53] to estimate the intrinsic and extrinsic camera parameters (calibration). Then we use a neural radiance field (NeRF-W [36]) to disentangle the static and transient components and obtain a dense 3D reconstruction (Figure 2, first block).

The 3D scene reconstruction offers rich 3D context - cameras and scene structure - enabling an in-depth study of humans (Figure 2, second block). First, for frames on the shot boundaries, the viewpoints from the two different shots act as effective multi-view (or multi-shot) information for human reconstruction [45]. We use the calibrated cameras and propose a multi-shot human reconstruction method, which jointly solves for body pose, body shape, identity and location. In this **calibrated multi-shot** method, camera information enables triangulation of people, which removes ambiguity and provides significant improvement

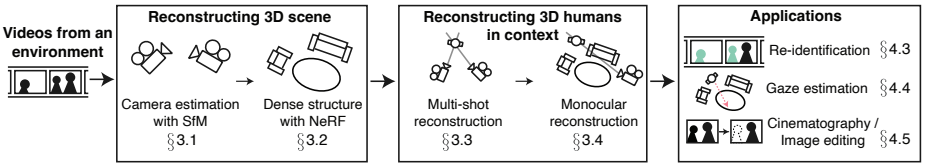

Fig. 2. Overview of our workflow. First, we use a collection of videos from a TV show environment and reconstruct the 3D scene (cameras and dense structure). We then use this information to recover accurate 3D pose and location of people over shot boundaries and on monocular frames. The recovered 3D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

upon the equivalent uncalibrated baseline [45]. Next, human reconstructions on the shot boundary inform us of the scale of each person relative to the scene. This is additional 3D context that is complementary to the cameras and scene structure. Since most frames are not on shot boundaries, we also formulate a monocular human reconstruction method that is explicitly guided by the extracted 3D context (camera, structure, body scale). The successful integration of the 3D context in our **contextual monocular** method leads to improvements over the state-of-the-art monocular baselines.

Our proposed "3Dification" of TV shows opens the door to many immediate applications (Figure 2, third block). First, our human reconstruction on the shot boundaries associates person detections, by incorporating geometric and anthropometric constraints. We show that this form of **re-identification** consistently outperforms traditional image-based baselines [12,22,23]. In parallel, from our reconstructed humans we can extract reliable **gaze information**, which can outperform specialized gaze estimation [46]. Moreover, our results provide estimates of the camera-to-person distance, which is relevant for **cinematography** applications [50,51]. Finally, we illustrate the potential use in **image editing** applications, like object insertion or human deletion.

In summary, our contributions can be summarized as follows:

- We identify the significant amount of 3D context (cameras, structure and body shape) in domains with repetition in the environment and the people, *e.g.*, TV shows, and propose a method to aggregate it from video sequences.
- We propose a formulation that integrates this context in 3D human estimation methods, which improves human reconstruction.
- We demonstrate how the aggregated 3D information can help a wide variety of downstream tasks: re-ID, gaze estimation, cinematography, image editing.
- We perform extensive qualitative and quantitative evaluation to validate the quality of our recovered 3D results.

2 Related work

2.1 Perceiving TV shows

The computer vision community has a long history of works on perceiving TV shows/movies. One of the most common tasks in this setting is studying the show characters with emphasis in face/character identification [2,9,39,43,57,58], where different cues have also been explored, *e.g.*, body, voice or gaze [6,8,34,35]. TV show data has been used extensively to study human behavior. Ferrari *et al.* study 2D pose estimation [10] and perform pose-based analysis [11]. Patron *et al.* [44] and Hoai *et al.* [19] focus on human interactions, while Recasens *et al.* [46] and Marín-Jiménez *et al.* [34] use this data to study gaze. Vondrick *et al.* [60] use sequences from TV shows to learn activity forecasting, while Wang *et al.* [61] leverage it for affordance learning. Despite this attention, all the above methods reason in 2D, with only a few exceptions. Everingham and Zisserman [9] use a 3D head model for re-identification. Here, we demonstrate how 3D location information can significantly simplify the re-ID problem. Pavlakos *et al.* [45] reconstruct humans from videos with multiple shots. However, they operate without camera calibration, while we show the importance of recovering reliable cameras.

2.2 Scene reconstruction

Reconstruction of 3D scenes is a well studied problem, *e.g.*, [1,21,53,54], however, most methods assume static scenes. Related work focuses on dynamic reconstruction [4,38], but requires capture from multiple wide-baseline synchronized cameras. Luo *et al.* [33] and Kopf *et al.* [29] present pipelines for recovering depth in monocular videos that include humans, but they assume that the underlying scene is static. View synthesis approaches like NeRF [37] and follow-ups [41,66] can be used to solve multi-view stereo, however these also assume static scenes. Other extensions of NeRF focus on reconstructing 3D motion in the scene [13,30,42], but are often limited when handling changes in appearances and transient objects. NeRF in the wild [36] is the most relevant approach for the type of data we use (Figure 3), since it can deal with appearance and transient changes. In this paper, we find that when our data is properly curated we can use NeRF-W to recover dense 3D structure. We then show that this structure can be used to guide consistent 3D human reconstruction.

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

2.3 Humans in 3D scenes

Most works that reconstruct humans in context with the scene assume a static, pre-captured 3D environment. Savva *et al.* [52] is one of the first works that explore 3D human-scene interactions from RGB-D video, while Hassan *et al.* [17] study the recovery of 3D humans in context with their environment from monocular images. Many works incorporate environmental constraints for motion estimation from videos [48,49,55,56,63,67,69], by assuming known floor or contact points. Recently, Guzon *et al.* [16] proposed a system for localizing a person in a known environment and estimating their 3D pose. Again, the environment is reconstructed *a priori* and the approach also requires an egocentric sensor and IMUs for pose estimation. Liu *et al.* [31] propose a method that reconstructs the scene and the people together using egocentric video captured in static outdoor scenes. In this work, we reconstruct structure from much more challenging dynamic scenes, by aggregating 3D information over video content.

Some works [62,68] have studied human reconstruction from single images, while also recovering aspects of the environment. PHOSA [68] recovers humans interacting with objects from in-the-wild images, and is followed by [62,64] in other settings. While they focus on visible human-object interactions, we consider cases where the scene might not be fully visible. Knowing camera parameters is an integral part of scene perception. SPEC [27] regresses camera parameters from a single image. In contrast, we can recover more reliable context for cameras by leveraging the whole collection of images from a TV show environment.

3 Technical approach

For the following discussion, we use the term *environment* to refer to a location, *i.e.*, a room, kitchen, cafe, etc., that appears often in a TV show. Figure 4 visualizes the panoramic view of the environments we reconstruct in this paper. We use the term *shot* for an uninterrupted sequence captured by a camera. Shots are organized in *scenes*, which are typically captured in the same environment. Multiple scenes comprise an *episode* and multiple episodes are organized into a *season*. In this work we collect videos across the whole season of a TV show.

3.1 Camera estimation

For the first step of our workflow, we need to register the cameras in a common coordinate frame (*i.e.*, computing intrinsics and extrinsics) for each environment. This amounts to hundreds of thousands of frames across the season. To keep the number of frames at a practical scale for Structure-from-Motion (SfM) pipelines, we sample frames at shot boundaries, which are automatically detected [23]. This helps to increase the variety of viewpoints - we only use two frames per shot, and inter-shot variety is typically larger than intra-shot variety.

On this reduced set of frames, we use DISK [59] to find correspondences. Since our data includes dynamic actors, we run Mask R-CNN [18] to detect human

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".

masks, and we reject correspondences on these regions. We use COLMAP [53] on the remaining feature matches and estimate the sparse 3D reconstruction and camera registration. We use a simple pinhole camera model, and allow each camera to have different focal length. For each frame t we get estimates of camera intrinsics $K_t \in \mathbb{R}^{3\times3}$ and extrinsics $R_t^{CW} \in \mathbb{R}^{3\times3}, T_t^{CW} \in \mathbb{R}^3$, where CW denotes camera to world transformation. This sparse reconstruction is used to register other frames (non shot-boundary images). Since we do not have access to 3D ground truth for TV show environments, the quality of our cameras is evaluated implicitly by the effect it has on the human reconstruction (see also Sup. Mat.).

3.2 Dense structure

Besides the camera registration returned from SfM, we also estimate the dense structure of the environment to help with human position estimation. Traditional dense reconstruction methods assume static scenes, but these assumptions are not satisfied in TV show environments which contain many images of extreme diversity (Figure 3). Instead, we use a NeRF-W network [36] for dense structure estimation. NeRF-W extends NeRF [37], to account for varying appearances and transient occluders. For efficiency, instead of training NeRF-W with all images, we use an automatic selection method to maximize viewpoint variety. We cluster the images based on camera location and viewing direction. For each cluster we select the image with least percent of Mask R-CNN human pixels to use for training (*i.e.*, maximum number of scene rays). After training, NeRF-W returns a volumetric 3D representation of the static structure of the scene.

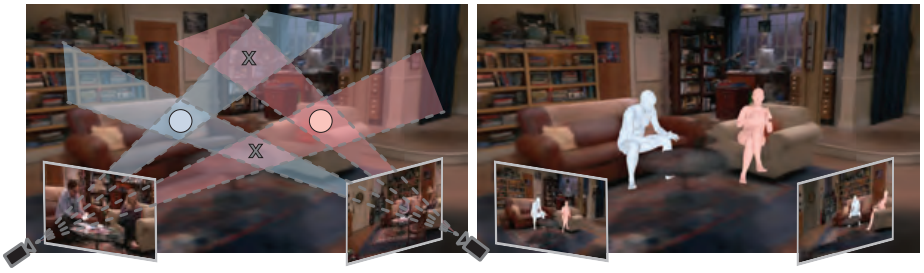

Fig. 5. Calibrated cameras for scale estimation and identity association. Given calibrated cameras, we can use frames at a shot change to solve for the actors' pose, location, relative scale and association. The four overlapping regions (left) indicate possible locations triangulated by the cameras.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

3.3 Calibrated multi-shot human reconstruction

In movies and TV shows, scenes are filmed in consecutive shots. The shot changes within a scene correspond to consecutive time frames seen by different viewpoints. This serves as *effective multi-view* information, providing signal to recover the 3D location and pose of the actors [45]. However, doing so requires knowledge of the identity of the actors across the shot changes. Prior work utilizes a pre-trained recognition-based re-ID model to establish these correspondences, but this is not always reliable, for example when only the back of the character is visible. We make an observation that when camera information is available, the association can be solved jointly with the 3D human pose, shape, and location. We refer to this approach as *calibrated multi-shot optimization*.

Let us assume there are M actors in frame t, N actors in frame $t+1$, and a shot change happens from frame t to $t+1$. We need to solve a matching problem to associate the two sets of actors. We propose to use the objective of SMPLify fitting [5] to model the cost for this matching.

Formally, let us consider a detection of a person at time instance t, with detected 2D keypoints $J_{est,t}$ [7]. We denote with θ_t the pose parameters and with β_t the shape parameters of the person in the SMPL format [32]. We use J_t for the joints and T_t^C for the translation of the body in the *camera frame*. Moreover, from SfM, we have access to the transformations R_t^{CW}, T_t^{CW} from the camera frame to the world frame at time t. Given all of the above, we minimize the objective function with respect to $\{\theta_t, \theta_{t+1}, \beta_t, \beta_{t+1}, T_t^C, T_{t+1}^C\}$:

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

Here, $E_{J_t} = E_{J_t}(\beta_t, \theta_t, K_t, J_{est,t})$ is the joints reprojection term and E_{priors} are anthropometric priors similar to [5]. The key constraint is multi-shot consistency, which encourages the estimated bodies to be similar in the global frame:

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

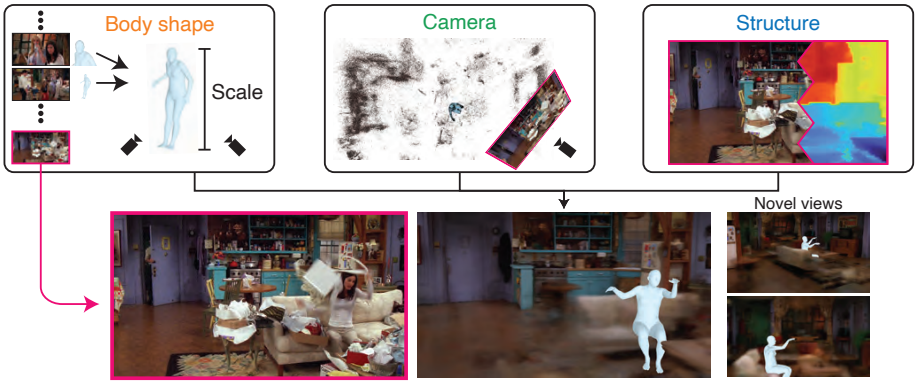

Fig. 6. Contextual monocular human reconstruction. For an input frame, we can leverage (a) the **body shape** (scale) of the person from a neighboring shot change, (b) the **camera** registration, and (c) the static **structure** of the environment. This enables monocular reconstruction of the person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

In contrast to prior work, we do not need to solve for the camera as we have access to reliable extrinsics and intrinsics (prior works [25,45,68] use a heuristic for focal lengths). This leads to more accurate human placement and constraints that allow for solving associations. Using this fitting cost E, we solve association by Hungarian matching. See Figure 5 for an illustration of this optimization.

3.4 Contextual monocular human reconstruction

Although shot changes provide effective multi-view information for free, the majority of the frames in the video only have monocular observations. Monocular human reconstruction is challenging, particularly so for TV shows with many close-up shots; however, in our case, we can capitalize on the contextual information we have recovered. In this subsection, we explain how we can make use of this 3D context in an effective way. We demonstrate this using a single-frame optimization approach, SMPLify [5], but other methods could also benefit from our context, *e.g.*, we show representative results for HuMoR [48] in the Sup. Mat.

A high-level overview of this step is presented in Figure 6. First, given the sparse reconstruction of the environment, we can register the **camera** for a new frame. This gives us both extrinsics R_t^{CW}, T_t^{CW} and intrinsics K_t for the camera via solving PnP with COLMAP [53]. We leverage these parameters for accurate projection. Moreover, we can employ the structure captured by our NeRF-W network. In general, it is not trivial to extract the structure from NeRF [41,66]. The native representation used by NeRF is in the form of densities for each point. Here, we propose to use this density as a proxy for occupancy of the 3D space. With this in mind, we formulate an objective to discourage the human

body vertices V from occupying areas with high density values:

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

where $\tilde{\sigma}$ samples values from the density field σ using trilinear interpolation, while ρ is the Geman-McClure robust error function [14]. Finally, we leverage the shape parameters $\hat{\beta}$ that capture the relative scale of the person with respect to the environment, and are recovered from the nearest shot change with the calibrated multi-shot reconstruction. This value can be used explicitly in the optimization to resolve the scale ambiguity.

Eventually, our monocular fitting objective minimizes:

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

with respect to θ_t, T_t^C, where we employ the **camera** information and the **body shape** parameters of the person during the fitting, while also discouraging the body mesh from penetrating the static **structure** of the scene.

3.5 Applications

An important argument in favor of 3D reconstruction for people in TV show environments is that it can simplify many reasoning tasks in this domain. For example, the calibrated multi-shot optimization explicitly reasons about the identity of the detected humans, as part of the Hungarian matching. This enables reliable *re-identification* in the challenging case of shot changes where the viewpoint can change significantly (Section 4.3). Moreover, one can extract *gaze information* from our 3D humans by considering the 3D pose of the face/head. With knowledge of camera pose, we can easily estimate the gaze direction in the global space, and thus compute gaze across the shot change (Section 4.4). Finally, we perform an analysis of our data which could be useful for *cinematography* applications, and highlight the potential of *image editing* using our results (Section 4.5).

4 Experiments

In this section, we present the quantitative and qualitative evaluation of our approach. We use seven popular TV shows (Figure 4) and one season from each. We follow the procedure described in Section 3 to collect the images we use. Each environment has 1k-5k frames from shot changes. For evaluation, we select per TV show a set of 50 person identities present on these shot changes. We use these frames as a test set to evaluate our method qualitatively with a crowd-sourced perceptual evaluation on AMT and curate it with the information we require for quantitative evaluation, *i.e.*, human-human associations, body keypoints, top-down location of the pelvis in the scene and gaze target across the shot change.

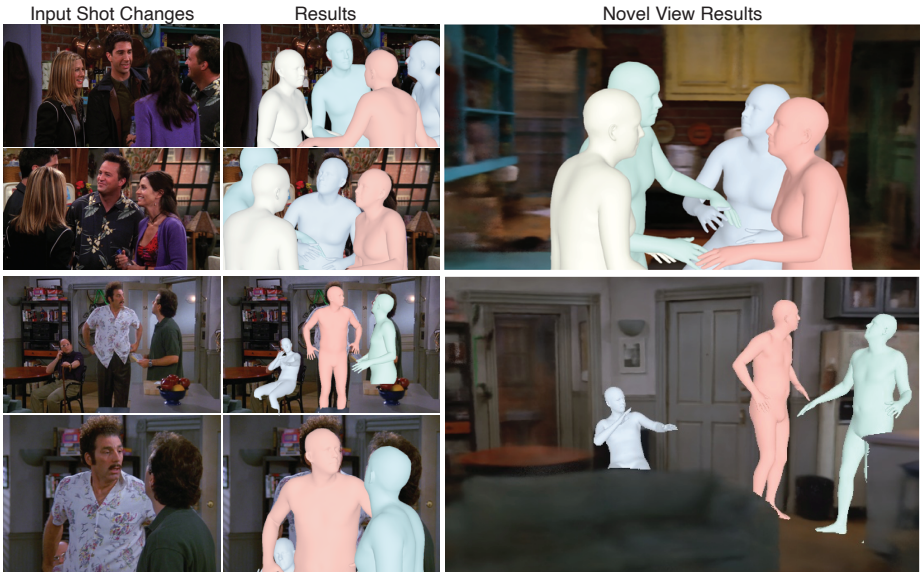

Fig. 7. Calibrated multi-shot and re-ID results. Using input shot changes (left), we perform our calibrated multi-shot optimization which jointly solves for pose, shape, location and association (middle). Note that identity, illustrated with colors, is not available a priori, but is estimated jointly with the 3D reconstruction. The recovered humans can be rendered in novel views using the NeRF of the environment (right).

4.1 Calibrated multi-shot human reconstruction

For a proof of concept, we first evaluate our proposed calibrated multi-shot optimization in a controlled setting, with the Human3.6M dataset [24], where we have accurate 3D ground truth for pose. Since our focus is on the effect of having access to camera parameters, we compare with the equivalent uncalibrated baseline, which is similar to [45]. The results are presented in Table 1. The significant improvement when having access to camera information further motivates the importance of our calibrated multi-shot algorithm.

For our data from TV shows, we do not have access to 3D ground truth for humans, so we perform two evaluations for the human reconstructions. First, we perform a system evaluation by Amazon Mechanical Turk (AMT) workers. For each 3D human reconstruction, we task the annotators to select the rendered result video (our method vs. a baseline) where the human reconstruction is more accurate and consistent with the scene and shot boundary images. Each result video is 10 seconds and provides multiple viewpoints of the person in the scene. We test on our test set, resulting in 2100 human labels from 48 participants who went through quality control (please see Sup. Mat. for more details). We report the percent of choices where our method is preferred over the baselines in Table 1. The uncalibrated baseline (first row; without intrinsics or extrinsics) is very rarely preferred over our calibrated baseline (last row; with estimated

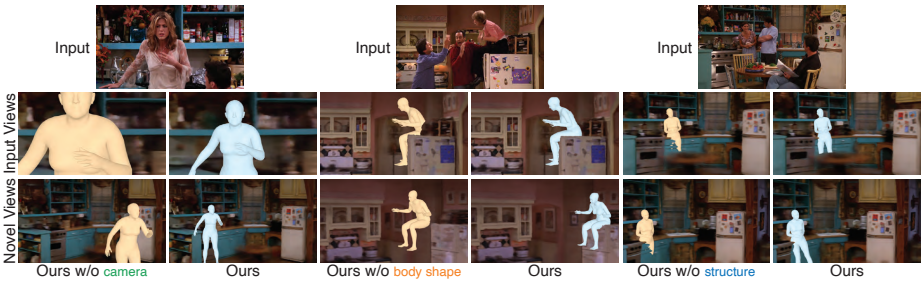

Fig. 8. Results for the contextual monocular reconstruction. We ablate the basic components of the contextual reconstruction to demonstrate their effects. Our method uses all three forms of context. Without our estimated camera intrinsics (left) and without body shape (middle), the person is incorrectly placed in the scene due to scale ambiguity. Using structure (right) avoids interpenetration with the environment.

intrinsics and extrinsics). Having access to estimated intrinsics can help with localization (middle row), but it is still preferred only 35% of the time. Besides the crowd-sourced evaluation, we also evaluate the location of each person quantitatively. In this case, we compute the mean metric distance error for the pelvis joint in the top-down projection, which is reported in Table 1. The conclusions are consistent with the AMT evaluation, highlighting the importance of camera information. Some representative results of our calibrated multi-shot optimization are presented in Figure 7, where we also indicate the estimated identity association (which we evaluate in more detail in Section 4.3).

4.2 Monocular contextual human reconstruction

Next, we investigate our proposed contextual monocular reconstruction. For this evaluation, we study the effect of each component separately – knowledge of the **cameras**, access to the person's **body shape**, and finally scene **structure** information. We present the results of this ablation in Table 2, where we report

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.

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.

cross-shot PCK @ $\alpha = 0.5$ [45]. Effectively, we project the person to the view across the shot boundary and measure localization accuracy for the joints in that space (more details in the Sup. Mat.). First, we see that state-of-the-art monocular methods without context [5,26,28] perform similarly on this data. Then, we examine the effect of context, using the optimization baseline [5] as our starting point (third row). Access to camera intrinsics is important to estimate a rough location of the person, and without it the method performs as the baseline without context. Knowledge of the body scale of the person, can make our estimate even more accurate. Finally, structure gives a smaller quantitative improvement but has a more pronounced qualitative effect by placing the person coherently in the environment. See Figure 8 for qualitative results.

4.3 Re-identification

For the re-ID evaluation, we examine the challenging case of person association after a shot change. For our case, re-ID is directly estimated from our calibrated

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.

multi-shot optimization. We compare this result with two types of baselines for computing affinities/costs between instances for Hungarian matching. The first type is image-based re-ID networks for affinity estimation, where [12] achieves SOTA on standard re-ID benchmarks, while [22,23] are trained on movies, a source of data similar to TV shows. The second type is a geometric baseline that uses our recovered cameras and is based on human keypoint triangulation, where the reprojection error is used as the cost for the Hungarian algorithm. Notice, that unlike SMPL fitting, this does not incorporate anthropometric constraints, *i.e.*, it considers every keypoint match independently, without using human body shape priors or measuring the holistic result. We report re-ID F1 scores in Table 3 using the visible pairs of actors before/after the shot change. Based on the results, our re-identification can consistently outperform these baselines.

4.4 Gaze estimation

For gaze estimation, we compare with the method of [46] that estimates the gaze target after the shot change. We evaluate the angular error in the gaze direction projected on the image plane. We report the Percentage of Correct Gaze Directions (similar to PCK [65]), using $\alpha = 20^o$ as threshold. Please see Sup. Mat. for details.

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].

Results are reported in Table 4. Since [46] relies on face detection, we report results on our whole test set (column "all") and on the subset where face detection is successful (column "w/ face"). Our approach outperforms [46] in both cases. Since we rely on body detection, we are more robust even when the face is occluded. Moreover, our extracted camera poses allow us to follow gaze across shots in a more accurate way. Further improvements are expected by modeling eye pose and saliency estimation to detect the gaze target, similarly to [46]

4.5 Cinematography/Image editing applications

We provide an initial analysis of our results in Figure 9, and present an extended study in the Sup. Mat. First, we visualize the *distribution of the estimated field of view for the cameras*. Here we can see the long-tail distribution for the views with a large field of view, *i.e.*, more informative viewpoints for 3D reconstruction. This justifies our insight to process data across the whole season, since the large majority of the views is typically close-ups. Moreover, we visualize the *locations of the cameras and the human actors*. The camera data could be useful for cinematography analysis [51], and the person data for behavior or affordance analysis [61]. Finally, we illustrate potential editing applications enabled by the

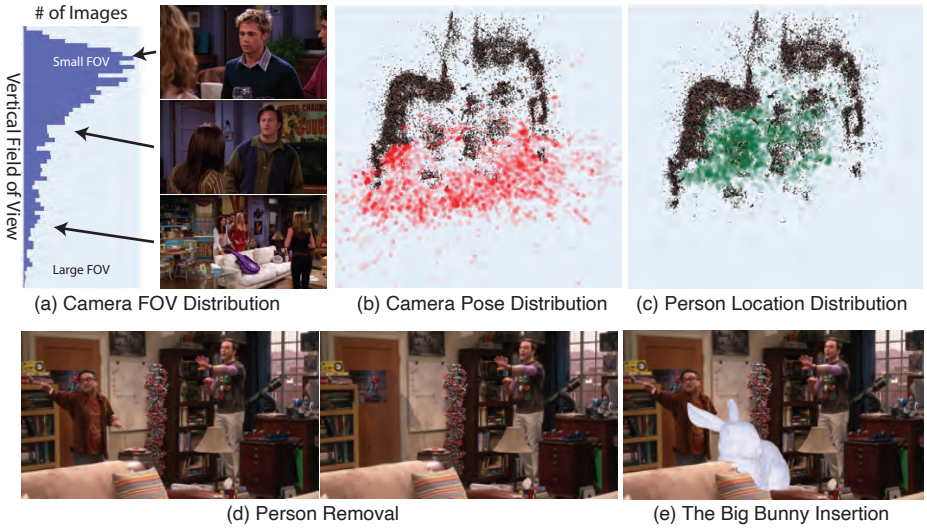

Fig. 9. Cinematography applications/Image editing. We present analysis of our processed data, including distribution of field of view, camera pose distribution and person location distribution for Friends (top). Moreover, we present editing options after our processing, including person removal and object insertion (bottom).

reconstruction of humans and the environment: *person removal and object insertion*. More editing options are possible, given the 3D nature of our processing.

5 Discussion

Conclusion: To the best of our knowledge, we are the first to reconstruct the people and the environment in TV shows and reason about them in 3D. We start with multi-shot video sequences associated with a specific environment and recover the camera, structure and relative human scale. We use this information as context to reconstruct humans even from a single frame, in a way that is consistent with their environment. We demonstrate our approach on seven different TV shows and present qualitative and quantitative results, as well as a wide variety of applications and analysis of the reconstructed data.

Future Directions: Our work has only scratched the surface of this extremely challenging and in-depth problem. Currently, we do not reconstruct the transient objects or dynamic objects that humans interact with (*e.g.*, chairs that move around, fridge opening). Also, the recovered pose of the humans is completely dependent on the quality of the 2D keypoint detections. It would be an interesting direction to incorporate appearance models for pose fitting.

Acknowledgements: This research was supported by the DARPA Machine Common Sense program as well as BAIR/BDD sponsors. Matthew Tancik is supported by the NSF GRFP.

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. In: ICCV (2009)
2. Arandjelovic, O., Zisserman, A.: Automatic face recognition for film character retrieval in feature-length films. In: CVPR (2005)
3. Arijon, D.: Grammar of the film language. Hastings House (1976)
4. Ballan, L., Brostow, G.J., Puwein, J., Pollefeys, M.: Unstructured video-based rendering: interactive exploration of casually captured videos. ACM Transactions on Graphics (TOG) **29**(4), 1–11 (2010)
5. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV (2016)
6. Brown, A., Kalogeiton, V., Zisserman, A.: Face, body, voice: Video person-clustering with multiple modalities. In: ICCVW (2021)
7. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. PAMI (2019)
8. Everingham, M., Sivic, J., Zisserman, A.: "Hello! My name is... Buffy" – Automatic naming of characters in TV video. In: BMVC (2006)
9. Everingham, M., Zisserman, A.: Identifying individuals in video by combining generative and discriminative head models. In: ICCV (2005)
10. Ferrari, V., Marín-Jiménez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR (2008)
11. Ferrari, V., Marín-Jiménez, M., Zisserman, A.: Pose search: retrieving people using their pose. In: CVPR (2009)
12. Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H., Chen, D.: Unsupervised pre-training for person re-identification. In: CVPR (2021)
13. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: ICCV (2021)
14. Geman, S., McClure, D.E.: Statistical methods for tomographic image reconstruction. Bulletin of the International Statistical Institute **4**, 5–21 (1987)
15. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: CVPR (2019)
16. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human POSEitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors. In: CVPR (2021)
17. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: ICCV (2019)
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
19. Hoai, M., Zisserman, A.: Talking heads: Detecting humans and recognizing their interactions. In: CVPR (2014)
20. Homayounfar, N., Fidler, S., Urtasun, R.: Sports field localization via deep structured models. In: CVPR (2017)
21. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: DeepMVS: Learning multi-view stereopsis. In: CVPR (2018)
22. Huang, Q., Liu, W., Lin, D.: Person search in videos with one portrait through visual and temporal links. In: ECCV (2018)
23. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: MovieNet: A holistic dataset for movie understanding. In: ECCV (2020)

24. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. PAMI (2013)
25. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: CVPR (2020)
26. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3D human body estimation. In: ICCV (2021)
27. Kocabas, M., Huang, C.H.P., Tesch, J., Muller, L., Hilliges, O., Black, M.J.: SPEC: Seeing people in the wild with an estimated camera. In: ICCV (2021)
28. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: ICCV (2021)
29. Kopf, J., Rong, X., Huang, J.B.: Robust consistent video depth estimation. In: CVPR (2021)
30. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: CVPR (2021)
31. Liu, M., Yang, D., Zhang, Y., Cui, Z., Rehg, J.M., Tang, S.: 4D human body capture from egocentric video via 3D scene grounding. In: 3DV (2021)
32. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 1–16 (2015)
33. Luo, X., Huang, J.B., Szeliski, R., Matzen, K., Kopf, J.: Consistent video depth estimation. ACM Transactions on Graphics (TOG) **39**(4), 71–1 (2020)
34. Marín-Jiménez, M.J., Kalogeiton, V., Medina-Suárez, P., Zisserman, A.: LAEO-Net++: Revisiting people looking at each other in videos. PAMI (2021)
35. Marín-Jiménez, M.J., Zisserman, A., Eichner, M., Ferrari, V.: Detecting people looking at each other in videos. IJCV (2014)
36. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: NeRF in the wild: Neural radiance fields for unconstrained photo collections. In: CVPR (2021)
37. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
38. Mustafa, A., Volino, M., Kim, H., Guillemaut, J.Y., Hilton, A.: Temporally coherent general dynamic scene reconstruction. IJCV (2021)
39. Nagrani, A., Zisserman, A.: From Benedict Cumberbatch to Sherlock Holmes: Character identification in TV series without a script. In: BMVC (2017)
40. Ng, E., Ginosar, S., Darrell, T., Joo, H.: Body2Hands: Learning to infer 3D hands from conversational gesture body dynamics. In: CVPR (2021)
41. Oechsle, M., Peng, S., Geiger, A.: UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: ICCV (2021)
42. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: ICCV (2021)
43. Parkhi, O.M., Rahtu, E., Cao, Q., Zisserman, A.: Automated video face labelling for films and TV material. PAMI (2018)
44. Patron-Perez, A., Marszalek, M., Reid, I., Zisserman, A.: Structured learning of human interactions in TV shows. PAMI (2012)
45. Pavlakos, G., Malik, J., Kanazawa, A.: Human mesh recovery from multiple shots. In: CVPR (2022)
46. Recasens, A., Vondrick, C., Khosla, A., Torralba, A.: Following gaze in video. In: ICCV (2017)

47. Rematas, K., Kemelmacher-Shlizerman, I., Curless, B., Seitz, S.: Soccer on your tabletop. In: CVPR (2018)
48. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: HuMoR: 3D human motion model for robust pose estimation. In: ICCV (2021)
49. Rempe, D., Guibas, L.J., Hertzmann, A., Russell, B., Villegas, R., Yang, J.: Contact and human dynamics from monocular video. In: ECCV (2020)
50. Savardi, M., Kovács, A.B., Signoroni, A., Benini, S.: CineScale: A dataset of cinematic shot scale in movies. Data in Brief **36**, 107002 (2021)
51. Savardi, M., Signoroni, A., Migliorati, P., Benini, S.: Shot scale analysis in movies by convolutional neural networks. In: ICIP (2018)
52. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: PiGraphs: learning interaction snapshots from observations. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016)
53. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
54. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: ECCV (2016)
55. Shimada, S., Golyanik, V., Xu, W., Pérez, P., Theobalt, C.: Neural monocular 3D human motion capture with physical awareness. ACM Transactions on Graphics (TOG) **40**(4), 1–15 (2021)
56. Shimada, S., Golyanik, V., Xu, W., Theobalt, C.: PhysCap: Physically plausible monocular 3D motion capture in real time. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020)
57. Sivic, J., Everingham, M., Zisserman, A.: "Who are you?" – Learning person specific classifiers from video. In: CVPR (2009)
58. Tapaswi, M., Law, M.T., Fidler, S.: Video face clustering with unknown number of clusters. In: ICCV (2019)
59. Tyszkiewicz, M.J., Fua, P., Trulls, E.: DISK: Learning local features with policy gradient. In: NeurIPS (2020)
60. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: CVPR (2016)
61. Wang, X., Girdhar, R., Gupta, A.: Binge watching: Scaling affordance learning from sitcoms. In: CVPR (2017)
62. Weng, Z., Yeung, S.: Holistic 3D human and scene mesh estimation from single view images. In: CVPR (2021)
63. Xie, K., Wang, T., Iqbal, U., Guo, Y., Fidler, S., Shkurti, F.: Physics-based human motion estimation and synthesis from videos. In: ICCV (2021)
64. Xu, X., Joo, H., Mori, G., Savva, M.: D3D-HOI: Dynamic 3D human-object interactions from videos. arXiv preprint arXiv:2108.08420 (2021)
65. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. PAMI (2012)
66. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: NeurIPS (2021)
67. Yuan, Y., Wei, S.E., Simon, T., Kitani, K., Saragih, J.: SimPoE: Simulated character control for 3D human pose estimation. In: CVPR (2021)
68. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3D human-object spatial arrangements from a single image in the wild. In: ECCV (2020)
69. Zhang, S., Zhang, Y., Bogo, F., Pollefeys, M., Tang, S.: Learning motion priors for 4D human body capture in 3D scenes. In: ICCV (2021)

70. Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K.G., Daniilidis, K.: MonoCap: Monocular human motion capture using a CNN coupled with a geometric prior. PAMI (2018)
71. Zhu, L., Rematas, K., Curless, B., Seitz, S.M., Kemelmacher-Shlizerman, I.: Reconstructing NBA players. In: ECCV (2020)