Supplementary Material for the paper: "The One Where They Reconstructed 3D Humans and Environments in TV Shows"

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

University of California, Berkeley

In this document we provide more details and analysis that were not included in the main paper due to space constraints. More specifically,

1. we discuss the results of the supplementary video (Section 1),
2. we include details about the captured/released data (Section 2),
3. we give statistics on camera/human distributions for our scenes (Section 3),
4. we discuss gaze estimation and evaluation (Section 4),
5. we include more qualitative results (Section 5),
6. we provide more details about the SfM optimization (Section 6),
7. we provide more details about NeRF-W training (Section 7),
8. we discuss the Amazon Mechanical Turk (AMT) evaluation (Section 8),
9. we present more experiments on the quality of estimated cameras (Section 9),
10. we discuss the importance of repetition in the data (Section 10),
11. we present experiments on temporal human reconstruction (Section 11),
12. we discuss our experimental evaluation (Section 12), and finally,
13. we include more details about the optimization procedures (Section 13).

We also encourage the readers to watch our supplementary video.

1 Results video

The included supplementary video provides extensive qualitative results from our 3D reconstruction of humans and scenes in TV show environments. We include results for a) our camera and structure recovery (Sections 3.1 and 3.2 of the main manuscript), b) our calibrated multi-shot human reconstruction (Section 3.3 of the main manuscript), and c) our contextual monocular human reconstruction (Section 3.4 of the main manuscript).

Moreover, although we follow the SMPLify paradigm [1] for these contextual results, as we discuss in the main manuscript (Section 3.4), we believe that our context can also inform other human reconstruction methods. To demonstrate this, we also provide results when using HuMoR [18] with our context, which leads to contextual temporal reconstructions. For this, we use the context we have estimated to inform the HuMoR optimization. This helps bypass some of the HuMoR assumptions (*e.g.*, static camera, fixed focal length), and demonstrates realistic contextual reconstructions. We hope that this will motivate other human reconstruction approaches to be applied on this type of data, and benefit from our context.

2 Data

For the data extraction, we use the official DVDs/BluRays for each TV show and for each season we study, we extract all the relevant frames. Although releasing the raw frames might not be possible due to copyright issues, we will release our frame extraction pipeline. Moreover, upon publication, we will release the results of our workflow, *i.e.*, a) SfM reconstructions, so that follow up work can register new images to the same coordinate frame, b) NeRF-W models, for view synthesis and structure estimation, c) our 3D human reconstructions, d) non-image information, *i.e.*, annotations and e) code for our approach.

3 Camera and human distribution analysis

In Section 4.5 of the main manuscript, we provided an analysis regarding the camera FoV/location distribution and the actors location distribution for the show "Friends". In this section, we extend this analysis by presenting the corresponding data for all the environments and TV shows that we study. Specifically, we present the statistics on image field of view distribution, camera location distributions, and people location distributions. The TV shows and specific locations/rooms investigated are shown in Table 1. The table is ordered from oldest to newest TV show, and all of the images at shot boundaries are considered for analysis in this section. In Figures 1, 2 and 3, we include the field of view (left) and spatial camera location distributions (middle) for all of these cameras at shot changes. Notice the long tail distribution where very few images have a large field of view. This is why it is important to consider many episodes, or in our case, a full season, to get a reliable reconstruction of the structure. On the right, we additionally show the spatial distribution of human locations. We believe these statistics can be interesting for additional observations/conclusions. For example, older TV shows tend to have more images with larger field of view, whereas for recent TV shows, the images with large field of view tend to be more

TV show	Environment	Season number	Number of images at shot boundaries	Number of images for training NeRF-W
Seinfeld	Jerry's apartment	9	4234	156
Friends	Monica's apartment	8	4196	167
Frasier	Crane's apartment	11	5287	165
Everybody Loves Raymond	Ray's apartment	9	5493	171
How I Met Your Mother	Ted's apartment	6	2310	121
Two And A Half Men	Alan's kitchen	10	3892	153
The Big Bang Theory	Sheldon's apartment	12	3808	165

Table 1. TV show environments and image information. We reconstruct seven TV show environments. For each environment, we consider an entire season of data. Each season contains many images, so we recover the camera parameters with SfM for the frames at the shot boundaries (fourth column). We cluster and filter these images into a smaller subset, which is used to train NeRF-W (fifth column).

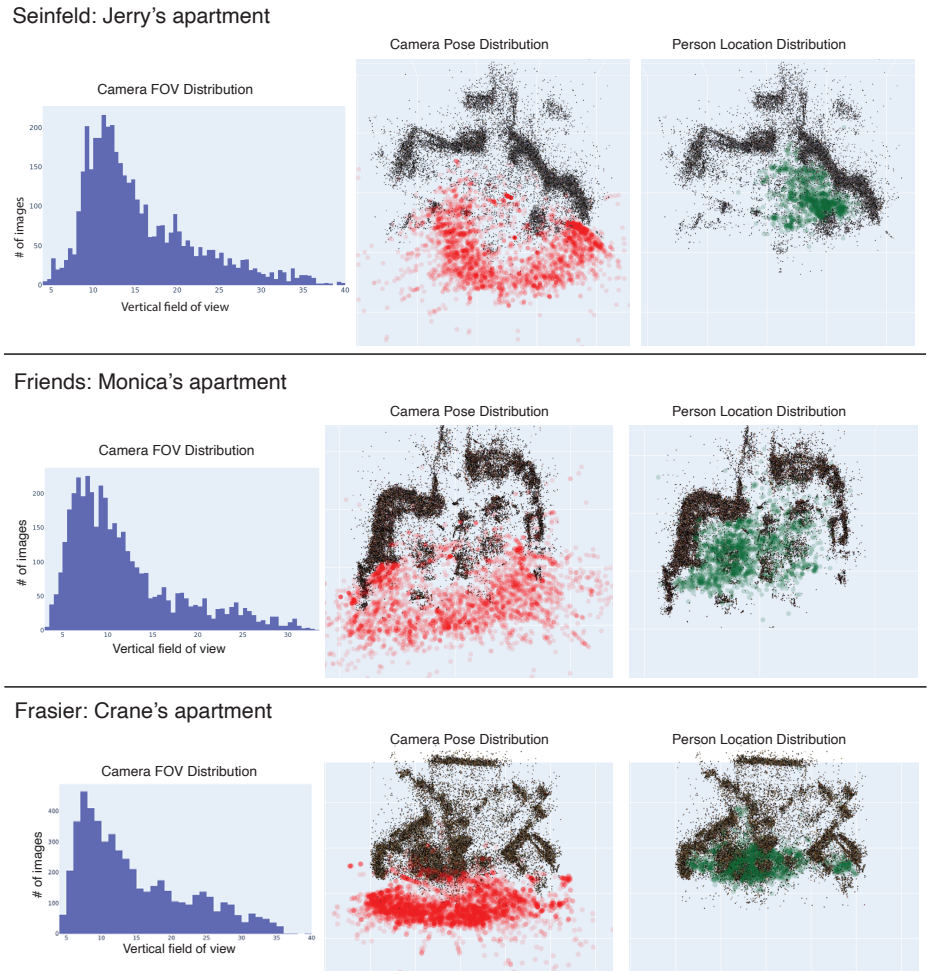

Fig. 1. Cameras and person location analysis (Part 1/3). Left: Camera field of view distribution. Notice the long tail distribution, where most images have small FoV–meaning the cameras are often zoomed in on the actors of interest. Middle: The camera pose distribution. Right: The person location distribution. Notice how some parts of the rooms are rarely traversed by people. These plots help to convey the data we are working with, in addition to providing insight into how TV shows are filmed with respect to both the camera and actor locations.

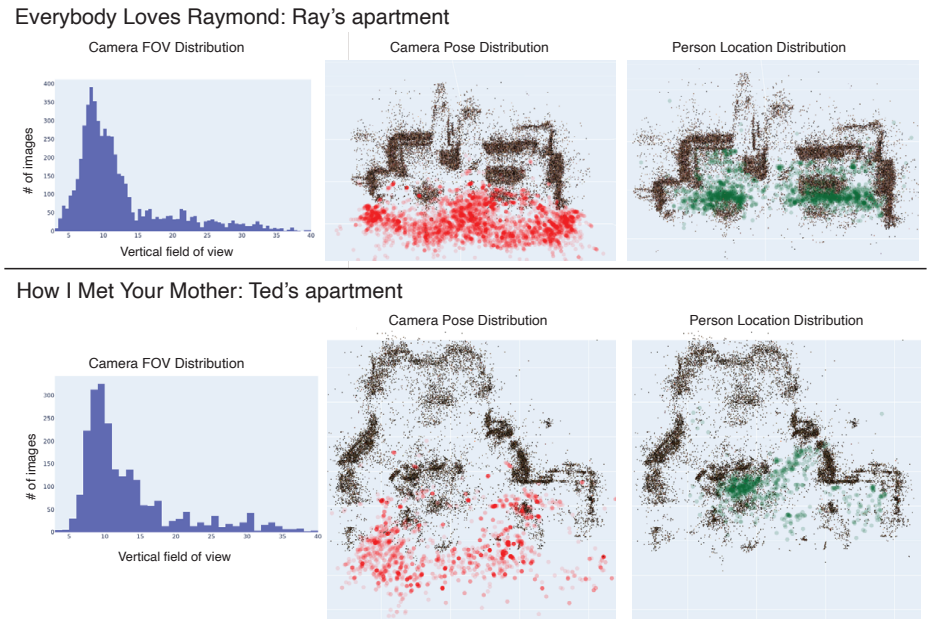

Fig. 2. Cameras and person location analysis (Part 2/3). Left: Camera field of view distribution. Notice the long tail distribution, where most images have small FoV–meaning the cameras are often zoomed in on the actors of interest. Middle: The camera pose distribution. Right: The person location distribution. Notice how some parts of the rooms are rarely traversed by people. These plots help to convey the data we are working with, in addition to providing insight into how TV shows are filmed with respect to both the camera and actor locations.

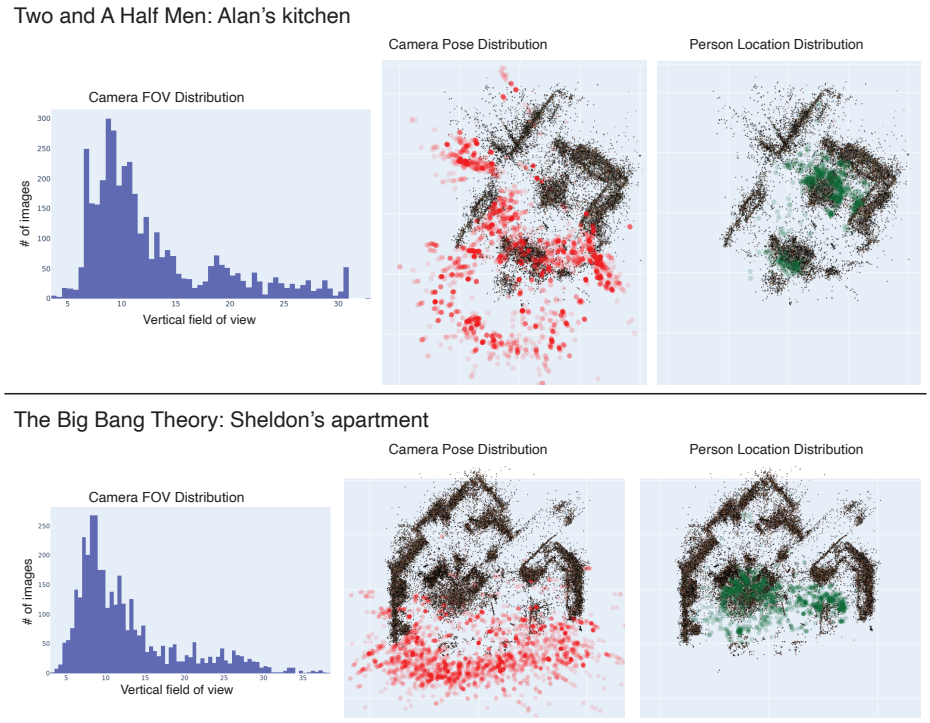

Fig. 3. Cameras and person location analysis (Part 3/3). Left: Camera field of view distribution. Notice the long tail distribution, where most images have small FoV–meaning the cameras are often zoomed in on the actors of interest. Middle: The camera pose distribution. Right: The person location distribution. Notice how some parts of the rooms are rarely traversed by people. These plots help to convey the data we are working with, in addition to providing insight into how TV shows are filmed with respect to both the camera and actor locations.

sparse. Moreover, some of the rooms are only occupied in certain parts of the layout, indicating the preference of the directors when filming the shows.

4 Gaze following

To estimate gaze given our 3D human reconstructions, we use a few selected vertices—one on the back of the head, and two on the eyes. Then, from a given mesh, the gaze direction is estimated as the ray passing through the center of the eyes and the vertex on the back of the head.

For gaze evaluation, we use the setting of Recasens *et al.* [17], which is the most relevant to our approach. We consider a person in the frame before the shot change, and we want to estimate the target of the person's gaze in the frame after the shot change. This requires estimating both the gaze direction of the person and the geometric transformation between the two frames. To simplify evaluation, because of the different formats of the two gazes (we return a full 3D direction, while [17] returns only a point), we compare gaze directions on the image plane (for the second image). As shown in the third image of Figure 4, we estimate a) the ground truth gaze direction (green line) based on the annotated gaze target, b) our estimated gaze direction (red line) based on the human reconstruction and c) the gaze direction for [17] (blue line) based on their estimated gaze target. Then, we estimate the error between the estimated and ground truth gaze direction, by computing the angular error between the two on the image plane (angle θ). Finally, we report results based on the percentage of correctly estimated gaze directions (PCGD), assuming a threshold α, where here $\alpha = 20^{\circ}$. In Figure 5 we provide more qualitative comparisons.

As we discuss in the main manuscript, the main two advantages we have compared to [17] is that a) we rely on the body of the person for detection (and reconstruction) instead of face detection as [17] does, which allows us to get a reasonable gaze estimate even for back or side facing people and b) we have a more explicit modeling of the 3D geometric relation between the two views. Further improvements for our results can be achieved by explicit modeling of the eye direction (*e.g.*, by using the SMPL-X model [14] which models the eyes), and by fusing the geometric gaze direction with image-based saliency methods for gaze target detection, similar to [17].

5 Additional qualitative evaluation

Figure 6 extends Figure 8 of the main manuscript, providing further examples that demonstrate the effect of the different factors in the contextual monocular reconstruction. In general, we observe that camera information (*i.e.*, focal length), is the most important cue, and helps the optimization to put the person in a reasonable location. This is important, because bad estimates for the focal length can significantly over/under-estimate the translation of the person in the space. Having a good estimate for the body shape helps to reduce the ambiguity in the scale during the reconstruction and tends to move the person in a more

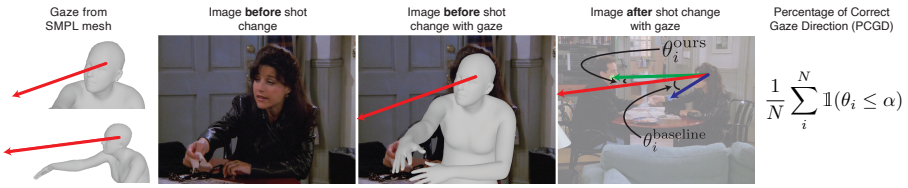

Fig. 4. Gaze following. We use the multi-shot reconstruction for the person on the shot boundary to estimate the gaze direction. This is defined by the ray passing through the center of the eyes and a specified vertex on the back of the head (first column). We visualize the frame before the shot change and the person reconstruction (second and third column). Given the relative pose between the two frames, we transform the gaze direction to the frame after the shot change (fourth column). The evaluation is done by comparing gaze directions on the image plane of the second frame. Using the annotated gaze target (green point) and the gaze target from [17] (blue point), we compute the corresponding gaze directions. We then compute the angular errors for each approach θ_i^{ours} and $\theta_i^{\text{baseline}}$. The reported error is based on the percentage of correct gaze directions for a specific threshold α, where here, $\alpha = 20^{\circ}$.

accurate position. Finally, using the static structure can help to avoid impossible configurations, *i.e.*, penetrating the walls (*e.g.*, moving out of the wall in rows 1 & 3, column 3 in Figure 6), the floor, the furniture (*e.g.*, moving out of the couch in rows 2 & 4, column 3 in Figure 6), or other surfaces. This helps fixing the final details when it comes to the location of the person.

6 Details on SfM optimization

Our workflow for estimating Structure-from-Motion on the environments we study is described in Section 3.1 of the main manuscript. Here we provide some more details of the SfM optimization. Specifically, we observed that COLMAP [19] can be sensitive to the initialization, so we perform the optimization in stages. After rejecting pixels that belong to humans (based on the output of Mask R-CNN [6]), we sort the images based on the number of non-human pixels. This sorting allows us to prioritize images with larger visibility of the background (*i.e.*, less zoomed in views), which will hopefully help with the stability during the SfM optimization. After sorting the images, we perform the optimization in stages. In the first stage, we start from scratch and use the *25%* of images with more valid (non-human) pixels. In the second stage, we initialize the optimization with the output registration from the first stage and use the *50%* of images with more valid pixels. In the third stage, we initialize the optimization with the output registration from the second stage and optimize over *all* images.

7 Details on NeRF-W training

In this section, we provide additional details for how we train a NeRF-W [12] network for the seven TV show environments considered.

Fig. 5. Qualitative results on gaze following. We present examples for the experiment on gaze following. For each example, in the first column, we present the image before the shot change. In the second column, we present our reconstruction and our gaze direction for the same image. In the third column, we present the gaze direction visualized on the image across the shot change. We include the ground truth (green), our result (red) and the baseline [17] (blue).

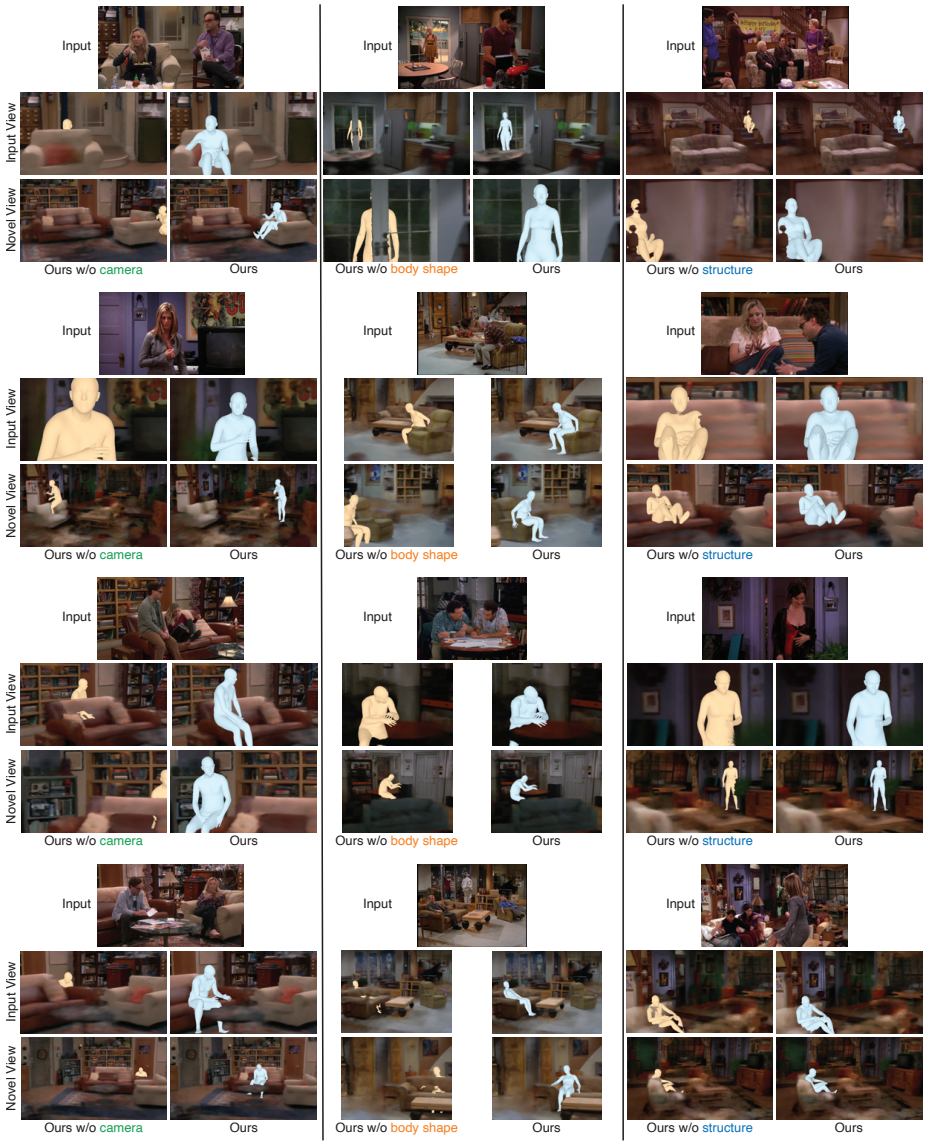

Fig. 6. Results for contextual monocular reconstruction. Here we show more results like Figure 8 of the paper. Recall that we ablate the basic components of the contextual monocular reconstruction to demonstrate their effects. Our method uses all three forms of context. Without our estimated camera intrinsics (left) and/or without body shape (middle), the person is incorrectly placed in the scene due to scale ambiguity. Using structure (right) avoids interpenetration with the environment.

Implementation details. We use the NeRF-W implementation from [15] with slight modifications. We train with 64 coarse samples and 64 fine samples per ray during training. Instead of defining a near and far plane per camera, we bound the ray samples by using a bounding box. More specifically, we create a 3D bounding box around the point clouds recovered by COLMAP [19] for each of the locations. At both training and test time, we compute the intersections of the ray with the box and only sample points within the box. We find that this modification helps to improve rendering quality, as samples are not wastefully sampled outside the scene. It also helps to avoid artifacts if rendering from camera locations outside the training set distribution. The main artifacts that remain happen for regions of the locations that are not observed, or are observed very rarely (*e.g.*, see supplementary video).

Choosing informative images. Each TV show environment has a few thousand images at the shot boundaries (Table 1) which COLMAP recovers the extrinsics and intrinsics for. The most straightforward way to train NeRF-W would be to use all of these images. However, we found that many of these images are not very informative for NeRF-W, since the majority of them have a small field-of-view as shown in Figures 1-3. Furthermore, because we consider images from an entire season of a TV show, many of the images have similar camera parameters and are thus redundant to the NeRF-W training process. In order to optimize for informative samples during the NeRF-W training process, we first discard images with a small field of view (smaller than $\sim 15°$) and then use a simple clustering method to optimize for large coverage of the scene with few images. We create a Plücker coordinate from each image by considering the camera origin and the ray going through the center pixel. We then use hierarchical clustering to create 200 clusters per scene, like those shown in Figure 7. We discard any clusters of size one because images in these clusters are likely to be outliers. Finally, from each cluster, we keep only the image with the least number of pixels detected as human from Mask R-CNN, to optimize for useful rays. Eventually, we are left with the number of images per TV show shown in Table 1 in the far right column.

Comparison to training with all images. As a simple comparison to demonstrate the effect of clustering, we show some qualitative results of training with all images compared to training with our filtered images in Figure 8. For the same number of rays seen during training, we observe that using the selected images produces less blurry renderings for RGB and more faithful structure for depth. We note that this comparison is informative in that training is more efficient and leads to cleaner results, but further analysis in this direction could be explored in the future. For example, one could investigate ways to detect and disregard images with significant transient effects, with differences in the generally static structure (*e.g.*, temporarily moved furniture, open doors), etc.

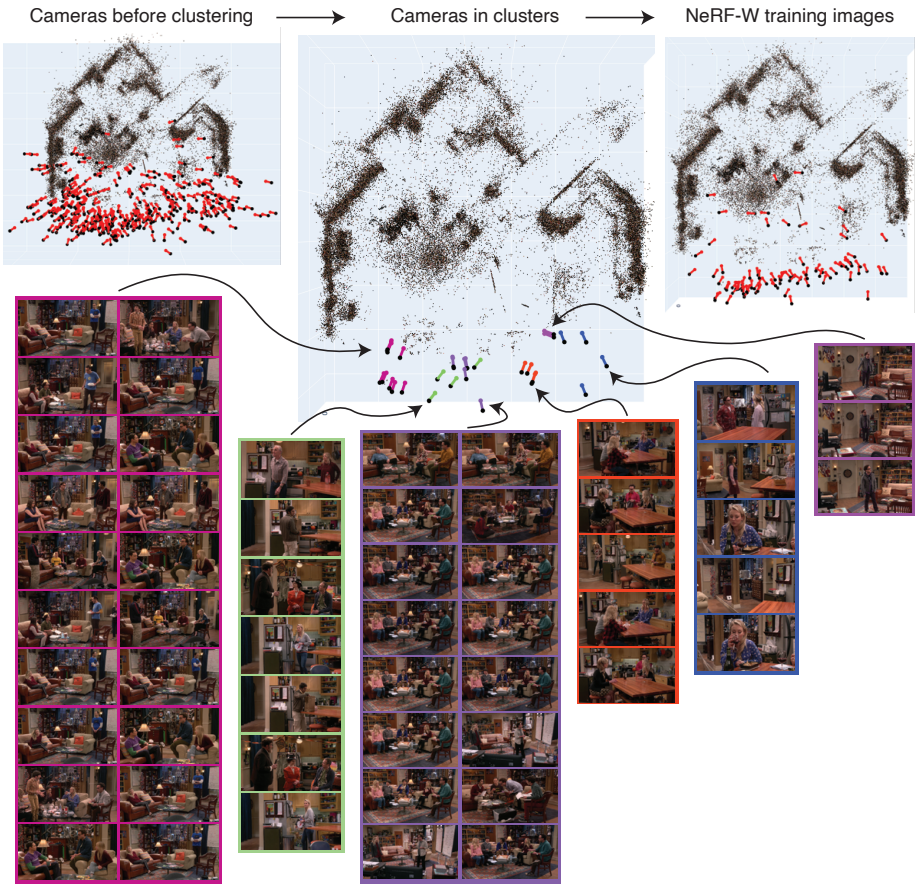

Fig. 7. Clustering and selecting images to train NeRF-W. We cluster the cameras recovered by SfM into clusters based on Plücker coordinates to find a smaller number of informative, diverse training images for training NeRF-W. (Top left) we show 10% of the COLMAP cameras and their viewing directions for Sheldon's apartment in The Big Bang Theory. (Middle and bottom) we show some of the 200 clusters that our pipeline creates. The colors indicate association with a particular cluster. (Top right) we show the 165 images after discarding clusters of size one and only keeping one informative image per cluster.

Fig. 8. NeRF-W with different training images. Here we show some static NeRF-W renderings on Sheldon's apartment of The Big Bang Theory. On the left, we show results for a model trained without any clustering or filtering. This model trains with the 3808 images at the shot boundaries. On the right, we show results for a model trained with our clustering and filtering procedure where 165 images are chosen. The far left images indicate where in the scene the rendering is occurring. We show both RGB and depth renderings, and we see that the model trained with our view selection has more desirable results.

8 Details on Amazon Mechanical Turk (AMT) evaluation

Code. Our code for the AMT evaluation will be released so others can run similar system evaluation experiments.

Evaluation data. The AMT task is used to qualitatively evaluate the effect of different calibration information in multi-shot human reconstruction. We provide the results in Table 1 of the main manuscript, where our proposed *calibrated multi-shot* version is at the bottom row. The AMT task consists of a force-choice question for 50 humans per each of our seven TV shows.

Question details. For each human reconstruction, we generate a 10 second video with a person-centric camera path that looks like a "bowtie" in order to show the human from multiple angles to expose any errors (*e.g.*, the human intersecting a couch, the human being in the wrong location in the scene, etc.). The videos are similar to those shown in the Sup. Mat., video but with custom per-person camera paths. For each question asked on AMT, we present the shot change images and bounding boxes overlayed on the person of interest. Below this, we show the two videos side-by-side with random ordering, and the user must choose the better video according to the question, "Which 3D visualization corresponds to where the person is in the room?".

Quality control. Each AMT task consists of 20 side-by-side video comparisons, but we hide additional videos in the task for **consistency** quality control. More specifically, we hide 3 consistency questions in the task. We also swap the order of the 3 hidden videos (*i.e.*, we change which video is shown first) and randomly place these questions in the task too. This results in 26 consecutive questions total per task. The annotators must respond consistently on all 3 hidden questions and their reversed versions in order to submit; otherwise, they will be notified of low consistency and their responses will be discarded.

9 Quality of estimated cameras

Given the nature of the TV show data (legacy content captured years or decades ago), it is not possible to get accurate 3D ground truth data regarding the camera information. Although we cannot explicitly evaluate the camera accuracy, we consider our results (*e.g.*, successful re-ID after the shot change, good gaze estimation) as an implicit indication of the reliable camera estimation.

To further validate the correctness of our recovered cameras from SfM, we perform an experiment by adding noise to the recovered camera parameters and reexamining the re-ID F1 scores (pointing to Table 3 of the main manuscript). More specifically, we perturb camera poses with random rotation and translation. We use a parameter n, and for various values of n, we apply a random rotation of n^o and a random translation with magnitude $n\%$ of the scene size (where scenes are roughly 10m). We then measure the effect of this noise on re-ID results after the calibrated multi-shot human reconstruction. As shown in Table 2, even small perturbations (*e.g.*, for $n = 1$, we have a rotation of 1^o and a translation < 10cm)

	No noise	n=1	n=2	n=5	n=10
Re-ID F1 on TV shows	**0.91**	0.79	0.64	0.42	0.33
MPJPE (mm) on Human3.6M [8]	**65.8**	72.9	90.3	158.0	269.3
PA-MPJPE (mm) on Human3.6M [8]	**47.1**	49.8	56.7	83.7	123.0

Table 2. The effect of adding noise to cameras on the calibrated multi-shot optimization. We add noise to the camera rotation and translation (different n values as described in text) and measure the effect on re-ID F1 scores for TV shows (top row) and human body pose accuracy, measured in MPJPE and PA-MPJPE (in mm) on Human3.6M (middle and bottom row).

immediately deteriorate performance. After $n = 2$, the re-ID F1 performance is worse than the image-based baselines from Table 3 of the main manuscript. This suggests that the originally recovered SfM cameras are high quality, otherwise they would not achieve competitive results for re-ID F1 scores.

We further examine the importance of accurate camera information for multi-shot human reconstruction, by using the setting of the Human3.6M dataset [8], where accurate 3D ground truth is available, and we can estimate the effect on pose estimation through the Mean Per Joint Position Error (MPJPE) and Procrustes Aligned MPJPE (PA-MPJPE) metrics [10,24]. For this evaluation, we synthesize shot changes by using consecutive frames in time, captured from different viewpoints. We use all actions from users S9 and S11, and we employ 2D keypoint detections from OpenPose [3] for the optimization. As we can see from Table 2, even small noise values can immediately affect the accuracy of the 3D pose result. This is further evidence that high quality cameras are important for accurate pose estimation from calibrated multi-shot reconstruction.

10 Importance of repetition in the data

The repetition in the data, *i.e.*, using images across a whole season of a TV show, is the key that allows us to recover the rich 3D context – good camera intrinsics and extrinsics from SfM, scene geometry and relative human scale. To highlight this even more, we perform an experiment where we ignore this repetition over the whole range of a season and only consider short sequences (∼5 secs) before and after each shot change for the SfM computation. This independent treatment of each sequence around the shot change leads to consistent failures in the COLMAP reconstruction. More specifically, for the shot changes included in our test set (Section 4 of main manuscript), for **45%** of the sequences, the reconstruction fails completely, for **42%** of the sequences, only a partial reconstruction is recovered (*i.e.*, it was not possible to register some cameras), while only for **13%** of the sequences, all cameras are registered in the same coordinate

frame. Given these common failures, without the data repetition, we would be unable to continue with the computation of scene structure & relative human scale, and extract the rich 3D context, which is crucial for our results.

11 Effect of context on temporal reconstruction

In the Supplementary Video, we provide qualitative results for temporal human reconstruction with and without the use of our recovered 3D context. For this demonstration, we employ the HuMoR [18] method for the temporal reconstruction. To better highlight this effect, we also present quantitative results using the cross-shot evaluation we employed for the contextual reconstruction. For this experiment, we employ PHALP [16] for tracking people in monocular sequences. For the most confident tracklets, we reconstruct their motions using HuMoR [18]. When we use HuMoR out of the box, the reconstruction only achieves a cross-shot PCK of **15.1%**. However, when we use our estimated cameras, the performance on the cross-shot PCK metric improves to **61%**. These results are not directly comparable with the numbers reported in Table 2 of the main manuscript, since some humans on the shot boundary might not be tracked successfully by PHALP, meaning that HuMoR can only operate on a subset of the test set. However, the improvement in performance for HuMoR is further quantitative indication of the effect of using our context for other methods too.

12 Experimental details

Data. For the experimental evaluation, we use frames captured at the shot boundaries. These frames correspond to sequential time instances, which means very similar 3D structure for the humans and the environment, while also being captured from different viewpoints, thus providing complementary information. We curate a set of 50 person instances for each of the seven TV shows. Each person is present on both the frame before and after the shot boundary, which translates to 700 appearances overall. This set of people and the corresponding frames are equipped with curated information for a) person identity, *i.e.*, information whether a person appears both before and after the shot change, b) 2D keypoints for the pose, c) location of the person (where we use the pelvis to specify the location of the person) in the scene from a top-down perspective, and d) gaze target on the image across the shot change. Identity is used in the person re-identification experiment, the keypoints are used in the contextual monocular reconstruction experiment, top-down location in the scene is used in the calibrated multi-shot human reconstruction experiments, while gaze target is used for the gaze following experiment.

Re-ID. For the person re-identification, we compare results from our calibrated multi-shot optimization with two types of baselines, one relying only on geometry, and one relying only on appearance (as always, further fusion of the

individual cues is feasible). For each baseline and our method, we compute affinity/matching costs, and then we do the matching using the Hungarian algorithm. We want to note that given a larger temporal window, one can use stronger baselines by clustering the identities [2,21], or building actor-specific models [20]. However, our goal is to highlight that in the simplest form of the problem, where only two frames are considered, using the geometry from registered cameras plus anthropometric information can lead to strong performance. Further improvements can be expected by fusing geometric & anthropometric information with image-based appearance information. Finally, it is important to note that this is also a very challenging setting, since we might only have very partial information for each person, *e.g.*, part of their back, which is challenging for methods that rely on face recognition.

Contextual monocular reconstruction. For the contextual monocular reconstruction results, we evaluate the effect of the context in monocular reconstruction. We use different baselines in the base setting without the use of context (Table 2 of the main manuscript) and then we show the benefits of context on an optimization similar to SMPLify [1]. Effectively, we reconstruct the person given one image, and then we project the 3D person on the image across the shot boundary, computing PCK on that image plane. This can be considered a proxy for 3D pose evaluation, since we evaluate the pose from a novel viewpoint. However, the focus of this evaluation is more about verifying the validity of the reconstruction in a holistic manner, and less about focusing on the detailed 3D pose. For a single-frame reconstruction to project well across the shot-boundary, we should have consistent cameras for the two viewpoints, a valid estimate of the body shape of the person from a neighboring shot change and a good estimation for structure that will be consistent with the human poses. In other words, this metric is more about holistic reconstruction, *i.e.*, recovering a body that is consistent with the accumulated context and less about capturing the nuances of the 3D human pose. If mm-level of accuracy is desired, one can use ground truth from MoCap, *e.g.* [5], but since MoCap is not possible to obtain for our in-the-wild TV setting, we believe that cross-shot PCK gives a good indication of the overall consistency of the result with respect to the context of the scene.

Calibrated multi-shot reconstruction. For the calibrated multi-shot reconstruction, we use the system evaluation by AMT workers to study the performance of our method. In presence of perfect cameras, we can expect very strong results from a calibrated multi-shot baseline, since this is very close to the calibrated multi-view setting, which has been considered in many occasions with ground truth cameras [4,7,9]. However, their setting is different, since we only rely on *estimates* for the camera intrinsics and extrinsics, instead of using ground truth values. Since we do not have access to accurate ground truth poses, we use this evaluation from AMT workers to assess the accuracy of the positioning of the people in the 3D space. This aspect is also evaluated quantitatively through the use of our pelvis position annotations from a top-down perspective in the 3D scene. To report the average distance errors with respect to these positions (Table 1 of the main manuscript), we rescale the scenes to a "common"/"average"

scale, since the SfM optimization cannot recover the absolute metric scale of the scene. Specifically, we consider all reconstructed humans from the calibrated multi-shot reconstruction and find their average height h_s for each scene s. We then use the overall average person height h_m, based on the SMPL model, and rescale the scene size by $\frac{h_m}{h_s}$. This effectively establishes that the people have roughly the same average height across all scenes, *i.e.*, we have a roughly "common" scale, even if we do not know the exact metric scale of each scene. We highlight that the position information we use for evaluation is important for many of the relevant works that study TV show data, *e.g.*, for affordance learning [23], human interactions [13] or activity forecasting [22]. For example, is the person sitting on the couch, or on the floor (see Figure 6 first row, first column), is the person inside or outside a room (see Figure 6 first row, second column)?

13 Optimization details

For the optimization, in the calibrated multi-shot setting, or the contextual single-frame setting, we start from an estimate of a regression network, similar to SPIN [11], but trained with cropping augmentation. After that, we follow with one optimization step for the single-frame reconstruction, or two for the calibrated multi-shot. In the case of multi-shot, in the first step we optimize over all parameters, and then optimize only over translation and body shape. The prior terms E_{priors} include the two body pose priors and the body shape prior from SMPLify [1], while the weights for the prior terms have the same values with [1]. In the multi-shot setting, we trust E_{glob} a lot, so it incurs a heavy penalty with a weight of $1e{+}6$. In the contextual monocular setting, $E_{\text{structure}}$ has a weight comparable to the other terms, 0.01.

References

1. 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)
2. Brown, A., Kalogeiton, V., Zisserman, A.: Face, body, voice: Video person-clustering with multiple modalities. In: ICCVW (2021)
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. PAMI (2019)
4. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3D pose estimation from multiple views. In: CVPR (2019)
5. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: ICCV (2019)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
7. Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M.J.: Towards accurate marker-less human shape and pose estimation over time. In: 3DV (2017)
8. 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)

9. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: ICCV (2019)
10. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
11. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: ICCV (2019)
12. 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)
13. Patron-Perez, A., Marszalek, M., Reid, I., Zisserman, A.: Structured learning of human interactions in TV shows. PAMI (2012)
14. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR (2019)
15. Quei-An, C.: Nerf_pl: a pytorch-lightning implementation of NeRF (2020), https://github.com/kwea123/nerf_pl/
16. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people by predicting 3D appearance, location and pose. In: CVPR (2022)
17. Recasens, A., Vondrick, C., Khosla, A., Torralba, A.: Following gaze in video. In: ICCV (2017)
18. 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)
19. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
20. Sivic, J., Everingham, M., Zisserman, A.: "Who are you?" – Learning person specific classifiers from video. In: CVPR (2009)
21. Tapaswi, M., Law, M.T., Fidler, S.: Video face clustering with unknown number of clusters. In: ICCV (2019)
22. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: CVPR (2016)
23. Wang, X., Girdhar, R., Gupta, A.: Binge watching: Scaling affordance learning from sitcoms. In: CVPR (2017)
24. 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)