# Human Mesh Recovery from Multiple Shots

Georgios Pavlakos, Jitendra Malik, Angjoo Kanazawa
University of California, Berkeley


Figure 1. **Human Mesh Recovery from Multiple Shots.** Videos from edited media, like movies, include sudden shot changes that lead to discontinuities between the frames (top), which reduce the rich potential of a film to a series of short independent temporal sequences. However, within the same scene, the underlying 4D structure of the scene changes smoothly. We leverage this insight and treat the different shots as multi-view cues that provide complementary information about the 3D human body underlying these shot boundaries. This leads to both more accurate 3D reconstructions (middle, bottom) and longer 3D pose sequences. These serve as a great source of data for training deep learning models that enable direct human mesh recovery on movie data.

## Abstract

*Videos from edited media like movies are a useful, yet under-explored source of information, with rich variety of appearance and interactions between humans depicted over a large temporal context. However, the richness of data comes at the expense of fundamental challenges such as abrupt shot changes and close up shots of actors with heavy truncation, which limits the applicability of existing 3D human understanding methods. In this paper, we address these limitations with the insight that while shot changes of the same scene incur a discontinuity between frames, the 3D structure of the scene still changes smoothly. This allows us to handle frames before and after the shot change as multi-view signal that provide strong cues to recover the 3D state of the actors. We propose a multi-shot optimization framework that realizes this insight, leading to improved 3D reconstruction and mining of sequences with pseudo-ground truth 3D human mesh. We treat this data as valuable supervision for models that enable human mesh recovery from movies; both from single image and from video, where we propose a transformer-based temporal encoder that can naturally handle missing observations due to shot changes in the input frames. We demonstrate the importance of our insight and proposed models through extensive experiments. The tools we develop open the door to processing and analyzing in 3D content from a large library of edited media, which could be helpful for many downstream applications. Code, models and data are available at:* https://geopavlakos.github.io/multishot/

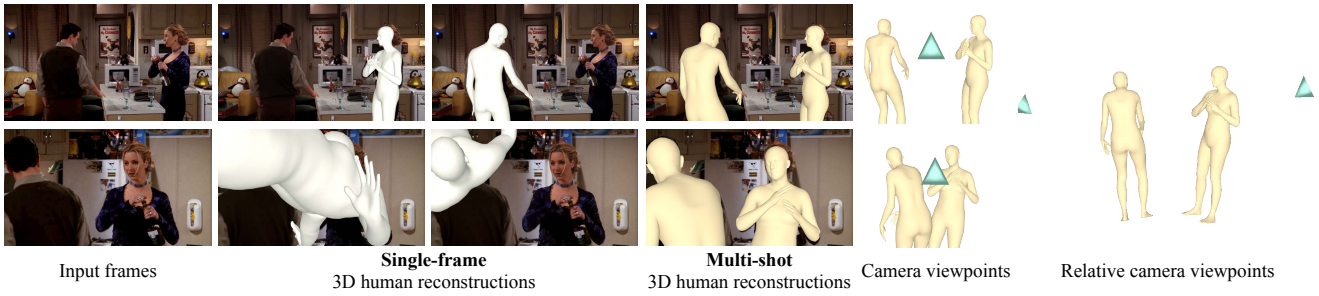


Figure 2. **Multi-shot reasoning.** Frames before and after the shot change depict the same 3D scene and provide a multi-view signal which helps reconstruct the underlying 3D pose of humans, particularly in cases of close-up, heavily truncated images of people. Blue triangles correspond to estimated camera locations in the scene. Each person is reconstructed independently.

## 1. Introduction

Movies are a treasure trove of human "behavior episodes" [4]. They are produced in many different countries in multiple genres, giving us tremendous cultural diversity and range. Datasets, most prominently, AVA [14] have emerged, which provide a rich annotation of spatio-temporally localized human actions in movies. This would seem like ideal data on which to train systems for video understanding, and furthermore use that as a stepping stone for acquiring "common sense" from observations of diverse human behavior. This "visual" route could be complementary to the "linguistic" route to capturing common sense and arguably more fundamental.

But before we go too far with our wishful thinking, we must confront a fundamental challenge of video data derived from movies – the complication of "shots". Film has a grammar [2]. Stories are communicated through a juxtaposition of shots, typically from different camera angles viewing the same scene. Alfred Hitchcock's Rope and Sam Mendes's 1917 are noteworthy precisely because they are presented as a single take, without any discernible breaks corresponding to shot boundaries.

These shot changes manifest as sudden discontinuities in video as illustrated in Figure 1. Current temporal 3D human mesh and motion recovery methods, as well as most action classification algorithms, treat these shots as independent scenes, which reduce the rich potential of a film to a series of short independent temporal sequences. Furthermore, shot changes often manifest in close up shots of actors and most state-of-the-art human mesh recovery models struggle to handle such heavily truncated images of people as shown in Figure 7. These two issues prevent applying such models to analyze 3D human behaviors in movies.

In this work, we propose a solution that addresses both of these challenges. First, we recognize that shot changes often depict a coherent underlying 4D scene from different viewpoints, despite the temporal discontinuities at the frame level. Thus, when handled properly, shot changes can be used as a multi-view signal of the underlying dynamic scene. This can be a powerful cue in disambiguating the 3D pose and motion of humans, which is particularly helpful for close-up, heavily truncated images of people (Figure 2). Specifically, we build on this novel and unexplored idea and propose a multi-shot optimization method that allows recovery of a consistent 3D human motion sequence *across* shot changes, simultaneously addressing both challenges of temporal fragmentation and partial humans.

The proposed multi-shot optimization allows recovery of long and reliable 3D human motion sequences from movies. This data can be treated as pseudo-ground truth and used for training regression models that predict human mesh directly from pixels in a feed-forward manner from images [22] or videos [23]. This workflow is illustrated in Figure 3. We show that high quality output from our multi-shot optimization is crucial for improving the performance of these models as multi-shot reasoning provides both longer and more accurate 3D pseudo-ground truth. Notably, unlike many previous works, the resulting direct prediction models are robust enough to perform human mesh recovery on movie data. Moreover, to further push the applicability on films, we propose a transformer-based architecture (t-HMMR) for our temporal encoder. A common challenge in edited media is that a person may not be consecutively depicted in the scene due to shot changes to another character or the background, often referred to as B-rolls (*e.g.*, sequence of Figure 3). Transformers can easily address this by explicitly not attending to frames that do not contain the person of interest and ignore them, while still processing a larger temporal context before and after the irrelevant input frames.

We experiment on AVA [14], a large scale dataset of movies with atomic action annotations. Applying our multi-shot optimization on AVA results in over 350k frames with pseudo-ground truth 3D. We treat this as training data to supervise regression models for human mesh recovery, from single image or video. Simultaneously, we curate a subset of AVA and use it for evaluation. We demonstrate the importance of our multi-shot optimization and the benefit on the downstream models through extensive experimentation

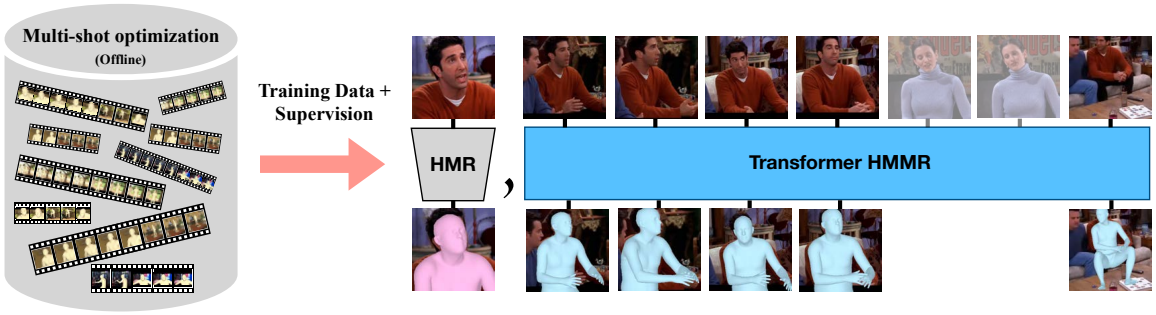


Figure 3. **Overview of our workflow.** We reconstruct 3D human mesh sequences from movies using our multi-shot optimization. The resulting reconstructions can be used as training data for both single-view human mesh recovery and temporal human mesh motion recovery.

on AVA and common benchmarks.

In summary, our contributions are:

- We introduce the problem of human mesh recovery from multiple shots and we propose an optimization approach that is applicable in multi-shot sequences. This results in high-quality 3D pseudo-ground truth that proves to be particularly effective at supervising direct regression models for human mesh recovery.
- We demonstrate that the resulting regression models can be applied successfully on movies, and we validate the importance of multi-shot reasoning at getting more accurate and longer pseudo-ground truth for training.
- To further push the applicability of regression models on movie data, we propose a temporal model with a pure transformer-based temporal encoder that is more suitable for inference on multi-shot sequences.

## 2. Background

This section provides reference to prior work and acts as background to our approach. The relevant literature is vast, so here we consider the most relevant approaches.

### 2.1. Human body modelling

Recent work in 3D human reconstruction has been influenced heavily by the availability of powerful human body models. The SMPL model [35] is one of the most popular choices that, among others, has enabled work on reconstruction [22], prediction [70], as well as imitation [46]. At a high level, one can consider SMPL as a function $\mathcal{M}(\theta, \beta)$ that takes as input pose parameters $\theta$ and shape parameters $\beta$ (collectively $\Theta = \{\theta, \beta\}$) and returns the 3D body mesh $M$ and joints $X$. Other body models follow similar formulations, with differences on the modelling side [42, 62, 64], or the expressivity of the model [1, 21, 43].

### 2.2. 3D pose and shape from single image

**Optimization:** Reconstructing 3D pose and shape from a single image is often addressed in an optimization setting. In these approaches [6, 15, 17, 30, 43, 68], a set of features are detected on an image (typically 2D keypoints), and then a configuration of the body model is recovered such that it is consistent with the features. This requires a reprojection objective $E_{\text{proj}}$ that penalizes deviations of the projected model from the detected features, and a set of objectives $E_{\text{prior}}$, that express the priors and encourage the reconstruction to be valid. At test time, the sum of these objectives is minimized in an iterative manner. The SMPLify [6, 43] methods are canonical examples of this type of approach for single image reconstruction, but other settings have also been considered, *e.g.*, from multiple views [10, 17] or monocular video [3, 24, 46, 50]. In this work, we adapt optimization approaches to be applicable in the setting of *multiple shots*.

**Direct prediction:** Directly regressing the SMPL parameters has seen many successes recently due to deep learning advances. A canonical example is HMR [22], which learns a direct mapping from raw RGB images to SMPL parameters and involves design principles adopted by many follow-up works [3, 13, 27, 44, 52]. More specifically, HMR consists of a feature encoder $f_{\text{im}} : I \mapsto \phi$ that converts an image $I$ to a feature representation $\phi$, followed by an iterative feedback regressor that maps the intermediate features to model parameters, $\hat{\Theta}$, and camera parameters, $\hat{\Pi}$. Using the predicted camera parameters, the reconstructed mesh can be projected to the image, which enables supervision with reprojection losses, given 2D annotations. Concurrently with HMR, other works have investigated decoupled regression approaches [9, 38, 41, 45, 56, 59, 65], where the intermediate feature representation is hardcoded, *e.g.*, 2D keypoints, silhouettes, semantic parts or dense correspondences. Recent works have introduced improvements to the HMR design, proposing camera estimation [26], probabilistic modelling [29, 53, 54], transformer-based architectures [33, 34], or other improved designs [25, 67, 69]. In this work, we adopt the HMR architecture for single-frame mesh recovery and following popular convention, we liberally refer to the model we use as HMR, even if the model weights are different than [22].

**Limitations:** Previous works [5, 20, 25, 51] have identified the limitations of relevant reconstruction approaches when it comes to heavy truncation of humans. Joo *et al*. [20] propose augmentation with synthetically cropped examples, Rockwell and Fouhey [51] retrain their model with confident reconstructions, while Kocabas *et al*. [25] propose a more robust architecture. In our work, we use complementary information from neighboring shots to improve the 3D reconstruction and collect training examples that improve the robustness of our single-frame model. Prior work has also identified the challenges and benefits of jointly reconstructing independent 3D instances, *e.g*., humans and humans [19, 40, 57], or humans and objects [63, 71]. Although we do not address these topics, we believe that multi-shot content could be helpful at perceiving these interactions.

### 2.3. 3D pose and shape from video

For video approaches, the goal is 3D reconstruction given a video sequence $V = \{I_t\}_{t=1}^{T}$, of length $T$. Video methods that follow-up HMR, *e.g*., [8, 23, 24, 36], take a similar workflow with the addition of a temporal encoder function $f_{\text{movie}}$, which maps per-frame features $\phi_t$ to per-frame sequence features $\Phi_t$, from which the model and camera parameters for each frame are predicted via a 3D regressor $f_{\text{3D}} : \Phi_t \mapsto \{\hat{\Theta}_t, \hat{\Pi}_t\}$. These methods differ in the choice of the architecture for the temporal encoder $f_{\text{movie}}$. Kanazawa *et al*. [23] use a convolutional model, Kocabas *et al*. [24], Choi *et al*. [8] and Luo *et al*. [36] use a recurrent model, while Sun *et al*. [58] use a hybrid model combining convolutions with self-attention. More recently, Rajasegaran *et al*. use a transformer architecture for spatio-temporal tracking [47] and temporal pose prediction [48]. In this work, we also investigate a pure transformer-based encoder, which is a more suitable architecture to handle missing identities that often occur in films.

### 2.4. Training with pseudo-ground truth

The strategy of using optimization approaches to generate pseudo-ground truth for human mesh regression models has been used in different contexts. For single images, Lassner *et al*. [30] use SMPLify [6] and manually discard failures to curate training data. SPIN [27] and EFT [20] build on this idea and initialize the optimization with an estimate provided by a regressor, which leads to more accurate fits, without requiring human intervention. Müller *et al*. [39] use a procedure similar to SPIN but focus on cases with self-contact. Arnab *et al*. [3] run a temporal optimization over monocular video, which can improve upon single frame results. Fang *et al*. [12] use mirror reflections as an additional view for resolving the depth ambiguity. Leroy *et al*. [31] focus on videos from the Mannequin Challenge [32], which provide multiple registered viewpoints in static scenes. In contrast to the above, in this work we investigate videos from edited media like movies, where many previous approaches are often failing and we capitalizing on the insight of multi-shot continuity to improve the quality of 3D pseudo-ground truth and the length of respective sequences.

## 3. Multi-shot optimization

Here we present the first step of our workflow based on multi-shot optimization. First, we describe the necessary preprocessing steps and the multi-shot optimization routine we use for pseudo-ground truth generation. Then, we provide more details about the application of our multi-shot optimization on the AVA dataset.

**Preprocessing** To apply our multi-shot optimization on a general video, we need a sequence of an individual within a scene.First, we detect 2D body joints using an off-the-shelf 2D pose tracker like OpenPose [7] or AlphaPose [11]. While these methods obtain quite reliable 2D joint tracklets, they fail across shot boundaries. To extend tracklet duration, we run a shot detection algorithm [49, 55], and use a person re-identification network trained on movie data [16] to link identities across shots. The result is longer 2D joint tracklets, extending beyond shot boundaries, which are used as inputs to the multi-shot optimization.

### 3.1. Multi-shot optimization

Relying on the insight that the input shots depict a single underlying 4D scene, we adapt optimization approaches such that they are applicable in the multi-shot setting and recover a consistent 3D human mesh across shot changes. To make this more concrete, let us consider the case where we have access to two consecutive frames $t$ and $t+1$, before and after the shot boundary respectively. As in SMPLify [6], we can setup data term $E^t_{\text{proj}}$ and prior term $E^t_{\text{prior}}$ for each frame. In order to incorporate the novel multi-shot insight, we introduce a term that encourages the body poses in these frames to be consistent. Note that prior works [3, 24, 46] have used temporal smoothness terms before, but we cannot naively apply these losses as done previously, because these approaches define smoothness regularization in the *camera coordinate frame*. This is because there is a large shot change in the camera frames due to shot changes.

As such, we must apply the smoothness regularization in the *canonical coordinate frame* in order to incorporate the multi-shot insight. Specifically, we explicitly decompose the pose parameters $\theta$ to global orientation $R_{\text{gl}}$ and body pose parameters $\theta_{\text{b}}$. By undoing the global orientation, we can compute the body joints $X_{\text{can}} = R^T_{\text{gl}} X$ in the canonical space. This formulation allows factoring out the camera motion, which can be abrupt, and imposing the smoothness

term only in the canonical frame:

$$E^t_{\text{sm joint}} = ||X^t_{\text{can}} - X^{t+1}_{\text{can}}||^2_2 \quad (1)$$

$$E^t_{\text{sm param}} = ||\theta^t_{\text{b}} - \theta^{t+1}_{\text{b}}||^2_2. \quad (2)$$

The sum of objectives is optimized over the entire sequence of length $T$:

$$E = \sum_{t=1}^{T}(E^t_{\text{proj}} + E^t_{\text{prior}}) + \sum_{t=1}^{T-1}(E^t_{\text{sm joint}} + E^t_{\text{sm param}}), \quad (3)$$

returning model parameters $\Theta^t$ for every frame $t$ of the sequence. For faster convergence to a more accurate solution, we initialize our reconstruction with pose and shape estimates provided by a regression network [27].

### 3.2. Reconstructing people in AVA

Although the above workflow is applicable in many occasions with videos from TV series or movies, in this work, we focus primarily on the AVA dataset [14]. AVA contains 300 movies annotated with human bounding boxes and atomic actions. Bounding box annotations are available at 1fps and organized in short tracklets. We also process the data at 1fps, and apply our preprocessing step to extend the tracklet duration over shot changes (*i.e.*, link short tracklets of the same identity). Each tracklet is reconstructed in 3D with our multi-shot optimization (section 3.1). Two important features of the reconstructed sequences are the diverse and challenging visual conditions (*e.g.*, truncation), and the length and quantity of the sequences that it includes. By reidentifying tracklets across shots, we can connect smaller, potentially overfragmented subsequences into longer multi-shot sequences, useful for training temporal models.

Our reconstructed sequences are treated as *pseudo*-ground truth. As is typical with relevant approaches that rely on pseudo-GT data sources [3, 23, 27], there might be errors in the detection of the 2D keypoints, in the tracklet re-ID, or the 3D reconstruction. Regardless, the quality of the pseudo-ground truth is demonstrated from the effect it has on the downstream task, *i.e.*, the training of deep learning models for human mesh recovery.

### 3.3. Evaluating 3D accuracy on AVA

Finally, our novel insight that pose changes smoothly across the shot boundary offers the opportunity to evaluate the 3D pose accuracy of the recovered human mesh on movie sequences without ground truth 3D data, via the concept of *novel view evaluation*. Specifically, for a shot change from frame $t$ to $t+1$, we project the mesh of frame $t$ to frame $t+1$, and vice versa. See Figure 4 and SupMat for details. This allows us to evaluate the predicted pose using 2D reprojection metrics, *e.g.*, PCK [66]. We refer to this metric as *cross-shot PCK* and use it to evaluate 3D pose quality in AVA, where 3D ground truth is not available.

To enable a more concrete evaluation on AVA, we manually curate AVA's test set. This curation includes human verification for tracklet re-ID, frames of shot changes and 2D keypoint locations, discarding examples where these steps are failing. All results reported on AVA refer to this clean subset, where we can reliably compute cross-shot PCK.

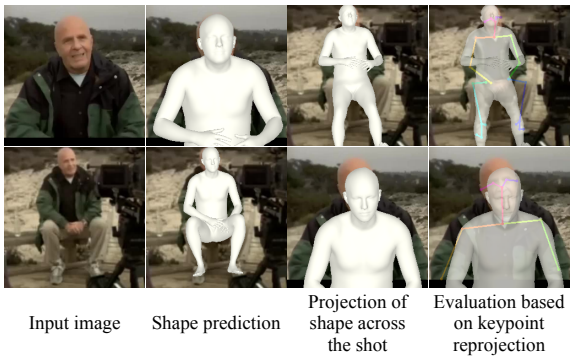


Figure 4. **Novel view evaluation with *cross-shot* PCK.** Given the shape prediction for frame $t$ (before the shot change), we project it to frame $t+1$ (after the shot change), and vice versa. We assess the 3D quality of the estimated pose by computing 2D reprojection metrics on this novel view.

## 4. Direct Human Mesh Recovery

The 3D motion sequences we recovered with the offline multi-shot optimization step offer a rich source of data with pseudo ground truth 3D bodies. Here, we demonstrate how to incorporate this data in the training of direct prediction models for Human Mesh Recovery from single images or video, without the reliance on keypoint detections.

### 4.1. Single-frame model

The first step is to train an updated single-frame model. In general, the setting is similar to the original HMR [22]. Let our image encoder for frame $I$ predict model parameters $\hat{\Theta}$ and camera parameters $\hat{\Pi}$. Model joints are projected to 2D locations $\hat{x}$. Our supervision for the network comes from the output of the multi-shot optimization for the corresponding frame, $\Theta_{\text{gt}}$, and the detected 2D joints $x_{\text{gt}}$.

$$L_{\text{2D}} = ||\hat{x} - x_{\text{gt}}||_1 \quad (4)$$

$$L_{\text{smpl}} = ||\hat{\Theta} - \Theta_{\text{gt}}||^2_2. \quad (5)$$

Our experiments show that training AVA dataset with our multi-shot 3D pseudo-ground truth improve the robustness of single-frame model against the diversity and the challenging visual conditions (*e.g.*, truncation).

### 4.2. Temporal model

Using an updated and robust single-frame model, we proceed towards learning the temporal encoding function

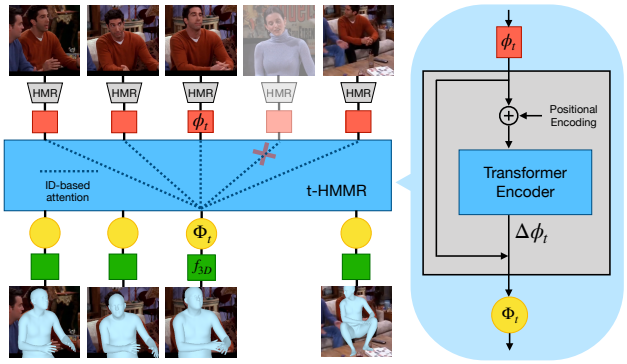


Figure 5. **Architecture of t-HMMR:** To most effectively leverage the plethora of 3D pose sequences recovered from our data, we propose t-HMMR, a human mesh and motion recovery model based on the transformer architecture. Even if the identity of interest is not present in some frames, we benefit from the large temporal context, by setting the attention to zero for the invalid frames, while aggregating information from the relevant input images.

$f_{\text{movie}}$. In the past, this function has been represented by convolutional [23], recurrent [24] or hybrid encoders [58]. However, all approaches assume a curated collection of clean videos of people with continuous person tracking [18, 37, 72]. In contrast, in more general use cases, including edited media, video data can be more challenging with issues like shot changes or B-rolls which interleave background frames in between shots. These cases are not easily handled by convolutional or recurrent encoders, which would require padding the inputs with zeros or concatenating all valid frames together, which ignores the difference in timestamps between concatenated frames. To address these limitations, we propose t-HMMR, a temporal model based on a pure transformer architecture [60]. Transformers include an attention mechanism, allowing us to explicitly select the elements of the input sequence they will attend to. This is a convenient feature, particularly with the discontinuous nature of sequences from films.

Our transformer encoder takes as input an intermediate feature embedding $\{\phi_t\}$ of sequence of frames $\{I_t\}$. This sequence comes with a scalar value per-frame $\{v_t\}$, which indicates whether the person is present in frame $t$ ($v_t = 1$), or not ($v_t = 0$). A fixed positional encoding $p_t$ is added to the input features to indicate the time instance $t$ of each input element. The updated features are then processed by a *transformer encoder layer*. This follows the architecture of the original transformer model, including a self-attention mechanism and a shallow feedforward network. The values $v_t$ are used to ensure that the invalid input frames will not contribute in the self-attention computation. The output of this layer is a residual value $\Delta\phi_t$ added to the feature $\phi_t$ through a residual connection. The final output is the video feature representation $\Phi_t$. This is illustrated in Figure 5.

For training the transformer encoder, following prior work [23, 24], we fix the weights of the image encoder $f_{\text{im}}$, and only update the temporal encoder $f_{\text{movie}}$ and the parameter regressor $f_{\text{3D}}$. Similarly to the single-frame model, supervision is provided by the multi-shot optimization results, where we have corresponding losses with Equations 4 and 5, $L^t_{\text{2D}}$ and $L^t_{\text{smpl}}$ respectively, for each frame $t$. Also, to further encourage temporal consistency, smoothness losses are applied on 3D joints $L^t_{\text{sm joint}}$ and 3D model parameters $L^t_{\text{sm joint}}$ (equivalent to equations 1 and 2 respectively).

## 5. Experiments

Our quantitative evaluation focuses on the effect of our multi-shot continuity insight in multiple aspects. First, we evaluate the efficacy of the multi-shot optimization; then we validate the quality of pseudo-ground truth provided from our offline multi-shot reconstruction by using it as supervision when training a single-frame human mesh recovery model; finally we also address temporal pose regression and highlight the importance of using multi-shot sequences for training, as well as employing a transformer-based architecture when dealing with movie data.

### 5.1. Experimental Setup

For single-frame regression, we use the HMR architecture [22] and adopt best practices from literature to establish a strong baseline: we train with the standard datasets using pseudo-ground truth SMPL parameters from SPIN [27], and use the recently proposed cropping augmentation scheme [21, 51]. We refer to this baseline as HMR$^+$ and use it for initialization of our multi-shot optimization and for ablative experiments. After the offline multi-shot optimization, our final single-frame model is trained with the same strategy, but with the addition of AVA dataset with pseudo-ground truth from our multi-shot optimization. We also compare with off-the-shelf baselines [22,25,27–29,51]. For the temporal model, we freeze the encoder of the single-frame model, as done in [23, 24], for computational efficiency, and train the temporal encoder and 3D regressor.

### 5.2. Multi-shot optimization

The proposed multi-shot optimization integrates information across the shot boundary to improve 3D pose reconstruction. To evaluate its success, we first setup a proof-of-concept experiment on Human3.6M [18], where 3D ground truth pose is available. Given the availability of multiple viewpoints, shot changes can be simulated by alternating camera views in the input sequence. We refer to the SupMat for more details on this evaluation. Then, we report results on AVA where we use the proposed cross-shot PCK metric (Section 3.2). With this evaluation, we investigate performance on the actual domain of interest (movies), while also providing additional quantitative validation with accu-

| Optimization | H3.6M (PA-MPJPE) ↓ | AVA (cross-shot PCK) ↑ |
|---|---|---|
| Single frame | 68.5 | 38.0 |
| Single shot | 62.7 | 42.3 |
| Multi shot | **59.2** | **55.2** |

Table 1. **Multi-shot optimization evaluation on Human3.6M and AVA.** We show PA-MPJPE (Human3.6M) and cross-shot PCK at $\alpha = 0.1$ (AVA). Our multi-shot optimization outperforms the optimization baselines applied on a single-frame or single-shot (temporal reasoning on frames that do not span shot changes)

rate 3D ground truth (Human3.6M). The results are presented in Table 1. As a sanity check, we compare with two optimization-based baselines, one that operates on a single frame [6], and one that operates on temporal sequence without shot changes [3,24,46]. In both cases, the multi-shot optimization outperforms the two baselines, which indicates that it can successfully integrate information across multiple shots. Qualitative examples of this behavior are presented in Figure 6 and in the SupMat.

### 5.3. Single-frame direct prediction

As described above, 3D pose sequences generated by our multi-shot optimization are used to supervise our direct regression models. Since the quality of the pseudo-ground truth affects the regression models, we can implicitly evaluate the importance of our multi-shot reasoning by investigating the effect it has on the downstream models. To achieve this, we present results on AVA, as well as PartialHumans [51] and 3DPW [61]. We provide ablations of our approach and comparisons against the most relevant state-of-the-art models [22,27,28,51]. For reference, we also report results from the most recent methods [25,29], although they might not be directly comparable to us (*e.g.*, stronger backbone & specialized architecture for [25]). Results are reported in Table 2, which lead to several insightful conclusions. *First*, on images from movies, many of the state-of the-art models perform poorly, and our pipeline allows us to improve performance on movie data compared to previous approaches. *Second*, we show that multi-shot optimization is a critical component in obtaining the best performance and naively training on AVA alone does not give as much improvements. Specifically, we conduct ablation studies where we train the base HMR$^+$ model with various AVA supervisions: 2D keypoints and pseudo-ground truth from single-frame optimization. We find that using the supervision from multi-shot reasoning achieves the best results. *Third*, the improvement we achieve from the supervision of multi-shot optimization is not specific to movie data only. Instead, we see improvement also on top of other challenging benchmarks; Partial Humans [51] and 3DPW [61]. Finally, we provide qualitative comparisons with the most relevant baselines in Figure 7, and include a discussion on failure cases in the SupMat.


Figure 6. **Qualitative effect of our multi-shot optimization.** Although a single frame baseline fails on the more challenging frames with heavy truncation (center), our multi-shot optimization leverages information from the less ambiguous frame across the shot boundary to get a more accurate 3D reconstruction (right).

| Method | AVA ↑ | PartialHumans ↑ | 3DPW ↓ |
|---|---|---|---|
| HMR [22] | 28.0 | 88.6 | 81.3 |
| GraphCMR [28] | 23.9 | 75.7 | 70.2 |
| SPIN [27] | 24.0 | 82.4 | 59.2 |
| Partial Humans* [51] | - | 83.3 | - |
| ProHMR [29] | 41.7 | 94.1 | 59.8 |
| PARE [25] | 40.8 | 94.4 | **50.9** |
| HMR$^+$ | 37.6 | 93.1 | 59.2 |
| + AVA (2D keypoints) | 32.0 | 93.9 | 58.5 |
| + AVA (single frame optim) | 41.1 | 93.9 | 59.3 |
| + AVA (multi shot optim) | **43.1** | **95.4** | 57.8 |

Table 2. **The importance of using pseudo-ground truth from multi-shot optimization when training a single-frame mesh recovery model.** We show cross-shot PCK at $\alpha = 0.1$ (AVA), PCKh (PartialHumans) and PA-MPJPE (3DPW). We compare our model trained with multi-shot pseudo-ground truth, with models trained with other forms of pseudo-ground truth (third block), as well as different state-of-the-art models (we share similar design with models in the first block; methods in the second block deviate from this). The availability of pseudo-ground truth from multi-shot optimization leads to improvements across the board.

### 5.4. Temporal model

The proposed multi-shot optimization not only provides better 3D pseudo-ground truth, but also has the benefit of mining long sequences to train temporal regression models on, as it can link sequences across shot changes. Here, we validate this and also evaluate the suitability of the proposed transformer temporal model, t-HMMR, on movie data.

Our analysis, performed on AVA, is summarized in Table 3. Again, we point to three interesting facts. *First*,


Figure 7. **Qualitative evaluation in the presence of truncations.** Comparison with the most relevant state-of-the-art on AVA [14] (first three rows) and Partial Humans dataset [51] (last two rows). Our model is significantly more robust in images with truncations.

we confirm that current state-of-the-art temporal models, HMMR [23], VIBE [24] and TCMR [8], when used off-the-shelf, have very low accuracy on movie sequences. *Second*, we observe that in the case of multi-shot movie sequences, the proposed transformer model outperforms other choices for the architecture of the encoder, i.e., convolutional [23] and recurrent [24]. As discussed, transformer can better handle missing identities (*e.g.*, due to b-rolls), which are common in edited media, and this translates also to a performance improvement.

*Finally*, we evaluate the performance gain coming from merging sequences from individual shots into a single sequence. For this, we use the exact same pseudo-ground truths from multi-shot, but split the sequences into individual shots (w/ single-shot AVA) and compare with the full model that is trained on merged sequences (w/ multi-shot AVA). Note that the only difference is the length of sequences used for training. Eventually, we identify performance improvement for the model when merging the individual shots into multi-shot sequences, which validates the importance of our multi-shot insight in mining longer sequences extending beyond a single shot.

In Figure 8 we provide example reconstructions of our t-HMMR model, in comparison with the single frame model, both trained on AVA. While the single frame model obtains reasonable results, output from t-HMMR is more consistent due to the larger temporal context.

| Model | same-frame PCK | cross-shot PCK |
|---|---|---|
| HMMR (Conv) [23] | 46.1 | 28.5 |
| VIBE (RNN) [24] | 40.1 | 25.0 |
| TCMR (RNN) [8] | 30.0 | 21.4 |
| Conv (w/ multi-shot AVA) | 79.6 | 53.6 |
| RNN (w/ mutli-shot AVA) | 78.3 | 52.6 |
| t-HMMR (w/ single-shot AVA) | 80.9 | 51.7 |
| t-HMMR (w/ multi-shot AVA) | **82.1** | **54.6** |

Table 3. **Multi-frame evaluation on AVA.** The numbers are same-frame and cross-shot PCK values. First three rows correspond to state-of-the-art models not trained on data from AVA. Using a) the transformer architecture, and b) our multi-shot insight to connect sequences that span more than one shots, is important to improve performance on movie sequences.


Figure 8. **Effect of temporal model.** While the single-frame prediction for frame 1 can be inconsistent with frame 2, our temporal model integrates information over the temporal window and estimates a body pose for frame 1 that is consistent with frame 2.

## 6. Conclusion

We introduce a new task of 3D human reconstruction from multiple shots. We propose an optimization approach, which in turn helps improving direct regression methods from single-frame and video. A limitation of the multi-shot reasoning is that it currently relies on Re-ID to identify which shots correspond to the same underlying scene and this can be noisy. Although our experiments show that even with this noise, the approaches benefit from multi-shot reasoning, it would be interesting to employ the most recent tracking systems [47, 48] to perform such re-identification. We believe that our work opens a new door towards analyzing movie data. In particular, our multi-shot reasoning provides relative extrinsic camera estimates between different shots. It would be exciting to use this information in the future to reconstruct not only the humans but also the rest of the environment. Movie data also exhibits "common sense" human behaviors that involve higher level reasoning. It would be interesting to analyze this in future work.

**Acknowledgements:** This research was supported by BAIR sponsors.

## References


[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Transactions on Graphics (TOG)*, 24(3):408–416, 2005. 3

[2] Daniel Arijon. Grammar of the film language. 1976. *Hastings House*, 1976. 2

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *CVPR*, 2019. 3, 4, 5, 7

[4] Roger G Barker and Herbert F Wright. Midwest and its children: The psychological ecology of an american town. 1955. 2

[5] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3D Multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data. *NeurIPS*, 2020. 4

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 3, 4, 7

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. *PAMI*, 2019. 4

[8] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *CVPR*, 2021. 4, 8

[9] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *ECCV*, 2020. 3

[10] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *ECCV*, 2020. 3

[11] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 4

[12] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3D human pose by watching humans in the mirror. In *CVPR*, 2021. 4

[13] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Kosecka, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, 2020. 3

[14] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 2, 5, 8

[15] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *ICCV*, 2009. 3

[16] Qingqiu Huang, Wentao Liu, and Dahua Lin. Person search in videos with one portrait through visual and temporal links. In *ECCV*, 2018. 4

[17] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, 2017. 3

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2013. 6

[19] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 4

[20] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *3DV*, 2021. 4

[21] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 3, 6

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 3, 5, 6, 7

[23] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 2, 4, 5, 6, 8

[24] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 3, 4, 6, 7, 8

[25] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 3, 4, 6, 7

[26] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. SPEC: Seeing people in the wild with an estimated camera. In *ICCV*, 2021. 3

[27] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 3, 4, 5, 6, 7

[28] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 6, 7

[29] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 3, 6, 7

[30] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 3, 4

[31] Vincent Leroy, Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, and Grégory Rogez. SMPLy benchmarking 3D human pose estimation in the wild. In *3DV*, 2020. 4

[32] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 4

[33] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 3

[34] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 3

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 3

[36] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3D human motion estimation via motion compression and refinement. In *ACCV*, 2020. 4

[37] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, 2017. 6

[38] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*, 2020. 3

[39] Lea Müller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *CVPR*, 2021. 4

[40] Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. Multi-person implicit reconstruction from a single image. In *CVPR*, 2021. 4

[41] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018. 3

[42] Ahmed AA Osman, Timo Bolkart, and Michael J Black. STAR: Sparse trained articulated human body regressor. In *ECCV*, 2020. 3

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 3

[44] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 3

[45] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 3

[46] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions on Graphics (TOG)*, 37(6):1–14, 2018. 3, 4, 7

[47] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3D representations. In *NeurIPS*, 2021. 4, 8

[48] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3D appearance, location & pose. In *CVPR*, 2022. 4, 8

[49] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *CVPR*, 2020. 4

[50] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. HuMoR: 3D human motion model for robust pose estimation. In *ICCV*, 2021. 3

[51] Chris Rockwell and David F Fouhey. Full-body awareness from partial observations. In *ECCV*, 2020. 4, 6, 7, 8

[52] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3D human recovery in the wild. In *ICCV*, 2019. 3

[53] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In *ICCV*, 2021. 3

[54] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In *CVPR*, 2021. 3

[55] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho, and Isabel Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177, 2011. 4

[56] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020. 3

[57] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *ICCV*, 2021. 4

[58] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, 2019. 4, 6

[59] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017. 3

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 6

[61] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018. 7

[62] Haoyang Wang, Riza Alp Güler, Iasonas Kokkinos, George Papandreou, and Stefanos Zafeiriou. BLSM: A bone-level skinned model of the human mesh. In *ECCV*, 2020. 3

[63] Zhenzhen Weng and Serena Yeung. Holistic 3D human and scene mesh estimation from single view images. In *CVPR*, 2021. 4

[64] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUMl: Generative 3D human shape and articulated pose models. In *CVPR*, 2020. 3

[65] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRac: Joint 3D pose and shape estimation by dense render-and-compare. In *ICCV*, 2019. 3

[66] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 35(12):2878–2890, 2012. 5

[67] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3D human pose and shape. In *CVPR*, 2021. 3

[68] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multi-

ple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, 2018. 3

[69] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, 2021. 3

[70] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3D human dynamics from video. In *ICCV*, 2019. 3

[71] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020. 4

[72] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 6