

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 only 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, which prove to be a great source of data for training deep learning models.

Abstract

Videos from edited media like movies are a useful, yet under-explored source of information. The rich variety of appearance and interactions between humans depicted over a large temporal context in these films could be a valuable source of data. 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 human 3D understanding methods. In this paper, we address these limitations with an 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, which leads to improved 3D reconstruction and mining of long sequences with pseudo ground truth 3D human mesh. We show that the resulting data is beneficial in the training of various human mesh recovery models: for single image, we achieve improved robustness; for video we propose a pure transformer-based temporal encoder, which can naturally handle missing observations due to shot changes in the input frames. We demonstrate the importance of the 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. Project page: <https://geopavlakos.github.io/multishot>

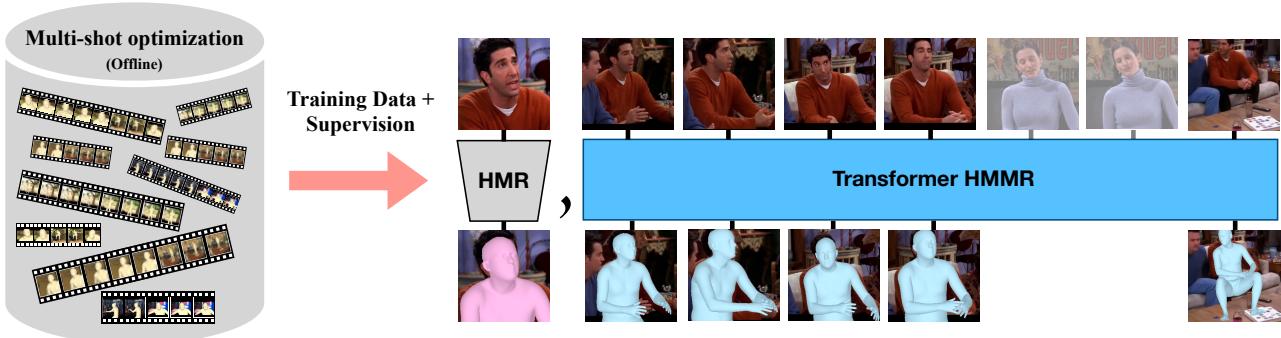


Figure 2: **Overview of our workflow.** We reconstruct 3D human mesh sequences from movies using multi-shot optimization. The resulting reconstructions can be used as training data for both single-view human mesh recovery and temporal human mesh motion recovery via the proposed transformer-based model.

1. Introduction

Movies are a treasure trove of human “behavior episodes” [5]. They are produced in many different countries in multiple genres, giving us tremendous cultural diversity and range. Datasets, most prominently, AVA [12] 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 [3]. 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. We see this as a lost opportunity, as these shots depict a single underlying 4D scene seen from different viewpoints despite the temporal discontinuities at the frame level. Properly modeled shot changes can thus be powerful signal rather than noise, as they provide a multi-view signal of the underlying dynamic scene. This can be a powerful cue in disambiguating the underlying the 3D pose and motion of humans, which is particularly helpful for close-up, heavily truncated image

of people. We take this insight and propose a multi-shot optimization procedure, which recovers a consistent 3D motion sequence across shot changes, simultaneously addressing both challenges of temporal fragmentation and partial humans.

The proposed multi-shot optimization recovers long 3D motion sequences from movies and serve as a rich source of pseudo-ground truth 3D training data for robust human mesh recovery from images or video. This workflow is illustrated in Figure 2. Training a single-frame human mesh recovery model [20] on the recovered 3D meshes from the multi-shot optimization results in improved robustness against heavy truncation and other ambiguities. More importantly, on the video setting, we propose a pure transformer-based temporal human mesh and motion recovery model (t-HMMR), which we demonstrate to be not only competitive, but also a particularly suitable architecture for films. Aside from sudden shot changes, films pose a challenge that the same person may not be consecutively depicted in the scene due to shot changes to another character or a background. Transformers can easily address these situations as it can explicitly not attend to frames that do not contain the person of interest and naturally ignore them, while still processing a larger temporal context before and after the missing input frames.

For our experiments, we employ AVA [12], a large scale dataset of movies with atomic action annotations. We apply our multi-shot optimization on AVA, which results in over 350k frames of data with pseudo ground truth 3D. We refer to this dataset as Multi-Shot-AVA (MS-AVA), and we use it to train regression models for human mesh recovery, both from single image (HMR) and from video (t-HMMR). We demonstrate the importance of our multi-shot optimization and the benefit of the collected data through extensive experimentation on MS-AVA and the common benchmarks.

In summary, we introduce the problem of human mesh

recovery from multiple shots and we propose a novel multi-shot optimization approach. This results in a new dataset, which is used to train a more robust single frame model for human mesh recovery and a pure transformer-based temporal model. Upon publication we will release our code and MS-AVA. We hope MS-AVA opens new opportunities for many future research directions.

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 [27] is one of the most popular choices that, among others, has enabled work on reconstruction [20], prediction [52], as well as imitation [35]. At a high level, one can consider SMPL as a function $\mathcal{M}(\theta, \beta)$ that takes as input pose parameters θ and shape parameters β (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 [47, 41, 48], or the expressivity of the model [2, 19, 32].

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, 13, 16, 25, 32, 51], 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_{proj} that penalizes deviations of the projected model from the detected features, and a set of objectives E_{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, 32] 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 [9, 16], or monocular video [4, 22]. In this work, we present an approach that focuses on the setting of reconstruction from *multiple shots*.

Direct prediction: Directly regressing the SMPL parameters has seen many successes recently due to deep learning advances. A canonical example is HMR [20], which learns a direct mapping from raw RGB images to SMPL parameters and involves design principles adopted by many follow-up works [4, 23, 11, 38, 33]. More specifically, HMR consist of a feature encoder $f_{\text{im}} : I \mapsto \phi$ that converts an image I to a feature representation ϕ , 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 [44, 8, 30, 31, 34, 42, 49], where the intermediate feature representation is hardcoded, *e.g.*, 2D keypoints, silhouettes, semantic parts or dense correspondences.

Limitations: Previous works [18, 37] have identified the limitations of relevant reconstruction approaches particularly when it comes to heavy truncation of humans. Joo *et al.* [18] propose augmentation with synthetically cropped examples, while Rockwell and Fouhey [37], retrain their model with confident reconstructions. 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 HMR model.

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.*, [21, 22, 28], take a similar workflow with the addition of a temporal encoder function f_{movie} , which maps per-frame features ϕ_t to per-frame sequence features Φ_t , from which the model and camera parameters for each frame are predicted via a 3D regressor $f_{3D} : \Phi_t \mapsto \{\hat{\Theta}_t, \hat{\Pi}_t\}$. These methods often differ in the choice of the architecture for the temporal encoder f_{movie} . Kanazawa *et al.* [21] use a convolutional model, Kocabas *et al.* [22] and Luo *et al.* [28] use a recurrent model, while Sun *et al.* [43] use a hybrid model combining convolutions with self-attention. In our work, we propose a pure transformer model, based exclusively on self-attention. We show that transformer encoder is not only a competitive and stable architecture for video, but also a suitable architecture to handle missing identities that often occur in films.

2.4. Common video datasets

Early work on 3D human reconstruction relied on motion capture datasets with full 3D supervision [40, 17] via mocap or multi-view stereo [29]. Penn Action [53] contains videos of people performing sport actions in front of a camera. 3DPW [46] is a recent video dataset with 3D ground truth of people outdoors obtained via IMUs. While these datasets are approaching “in the wild” settings outside a motion capture lab, they are still designed for performance capture where viewpoints are biased towards those that depict the entire body of people. This is in contrast to how people are depicted in films [12, 15] and other edited media [37], which are abound with shot changes and close-up shots of people that induce a large truncation. We propose both optimization and direct prediction methods that can address these challenges and introduce MS-AVA in goals of

further advancing 3D human mesh recovery from video.

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 MS-AVA dataset we introduce.

3.1. Preprocessing steps

To apply our multi-shot optimization on a general video, we need a sequence of an individual within a scene, including shot changes. First, we detect 2D body joints using an off-the-shelf 2D pose tracker like OpenPose [7] or Alpha-Pose [10]. While these methods obtain quite reliable 2D joint tracklets, they fail across shot boundaries. To extend tracklet duration, we run a shot detection algorithm [39, 36], and use a person re-identification network trained on movie data [14] to link identities across shots. The result is long 2D joint tracklets, extending beyond shot boundaries, which are used as inputs to the multi-shot optimization.

3.2. Multi-shot optimization

Relying on the insight that the input shots depict a single underlying 4D scene, we propose a multi-shot optimization method that recovers a reliable 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. In the SMPLify [6] setting, we have data terms E_{proj}^t and prior terms E_{prior}^t for both frames. In the case of sequence input, relevant methods [4, 22] include smoothness regularization on the *camera frame*. Since the shot (and therefore the viewpoint) has changed, this is not applicable here. As a result, these methods would consider the shot change as a discontinuity and reconstruct each frame independently.

In contrast to that, to leverage the continuity of the 3D body structure, we apply smoothness regularization in the *canonical frame*. More specifically, we explicitly decompose the pose parameters θ to global orientation R_{gl} and body pose parameters θ_b . By undoing the global orientation, we can compute the body joints $X_{\text{can}} = R_{\text{gl}}^T X$ 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:

$$E_{\text{sm joint}}^t = \|X_{\text{can}}^t - X_{\text{can}}^{t+1}\|_2^2 \quad (1)$$

$$E_{\text{sm param}}^t = \|\theta_b^t - \theta_b^{t+1}\|_2^2. \quad (2)$$

The sum of objectives is optimized over the entire sequence

of length T :

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

returning model parameters Θ^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 [23].

3.3. Multi-Shot AVA dataset

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 [12]. 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, but follow our preprocessing to extend the tracklet duration (i.e., link short tracklet of the same identity). Each tracklet is reconstructed in 3D with our multi-shot optimization (section 3.2). We call this new dataset **Multi Shot AVA**, or **MS-AVA**.

Two important features of MS-AVA are the diverse and challenging visual conditions (*e.g.* truncation), and the length and quantity of the sequences that it includes. In contrast to previous datasets [17, 46, 53], with MS-AVA, we are able to recover *many, long* and *diverse* tracklets. There are two factors that help us achieve that: a) connecting tracklets across shots and b) keeping track of person identities, even if they are missing for some frames. Both aspects allow us to connect smaller, potentially overfragmented subsequences into longer sequences, useful for training temporal models. This statistic is also highlighted in Table 1. More specifically, compared to single shot processing (“AVA single-shot”), we can increase the number of longer tracklets, if we consider identities that appear consecutively across shots (“AVA continuous identity”), or by re-identification after absence in some frames (“MS-AVA”). Examples of such sequences are illustrated in Figure 3.

Finally, our insight that pose changes smoothly across the shot boundary also offers the opportunity to evaluate the 3D accuracy of the recovered human mesh from monocular camera sequences via *novel view evaluation*. Specifically, given a shot change from frame t to frame $t + 1$, we project the shape of frame t to frame $t + 1$, and vice versa. See Figure 4 for illustration. This allows us to evaluate the predicted shape using 2D reprojection metrics, *e.g.*, PCK [50]. We refer to this metric as *cross-shot PCK* and we use it to evaluate 3D shape quality in AVA, where only 2D keypoints are present.

Dataset	Length (hours)		#Tracklets
	All / Long	All / Long	
Human3.6M [17]	6.45 / 6.45	330 / 330	
3DPW [46]	0.68 / 0.56	87 / 58	
Penn Action [53]	0.85 / 0	2.3k / 0	
AVA single-shot	75.1 / 11.1	5.1k / 1.3k	
AVA continuous identity	78.4 / 17.1	6.3k / 1.8k	
MS-AVA	81.1 / 20.2	6.7k / 2.1k	

Table 1: Raw duration and sequences for video datasets Numbers are for a) all tracklets and b) long ones (length greater than 20 secs). MS-AVA provides many long sequences which is supported by considering multi-shot continuity and re-identification across disjoint subsequences (see also Figure 3).



Figure 3: Different type of sequences in MS-AVA. The top row illustrates the most common type of sequence used in prior work with continuous sequence of an individual without shot changes. Movies include shot changes (middle) and show changes with non-continuous identity (bottom). Our approach can handle all three settings.

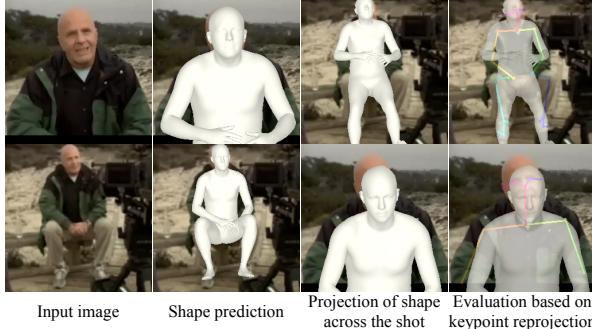


Figure 4: Novel view evaluation with cross-shot PCK. Given the shape prediction for frame t (before the shot change), we project it to frame $t+1$ (after the shot change), and vice versa. We assess the 3D quality of the estimated shape by computing 2D reprojection metrics on this novel view.

4. 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 [20]. 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, Θ_{gt} , and the detected 2D joints x_{gt} .

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

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

Our experiments show that the diversity and the challenging visual conditions (*e.g.*, truncation) of MS-AVA, help improve the robustness of our single-frame model.

4.2. Temporal model

Using an updated and robust HMR, we proceed towards learning the temporal encoding function f_{movie} . In the past, this function has been represented by convolutional [21], recurrent [22] or hybrid encoders [43]. However, in the above cases, the temporal training data come from curated collections of clean videos with continuous person tracking [17, 29, 53]. In contrast, in more general use cases, including MS-AVA, video data can be more challenging with issues like shot changes or frames where the person of interest is not present (due to occlusion, tracking failures or re-identification failures). 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 (*i.e.*, ignoring the time difference between consecutive frames).

To address these limitations, we propose t-HMMR, a temporal model based on a pure transformer architecture [45]. 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 MS-AVA sequences (see Figure 3). Only considering the sequences where a person appears continuously without identity changes, limits the available training sequences (*e.g.*, Table 1 for MS-AVA). The proposed transformer model has the advantage to address this challenge elegantly, making effective use of all the available sequences.

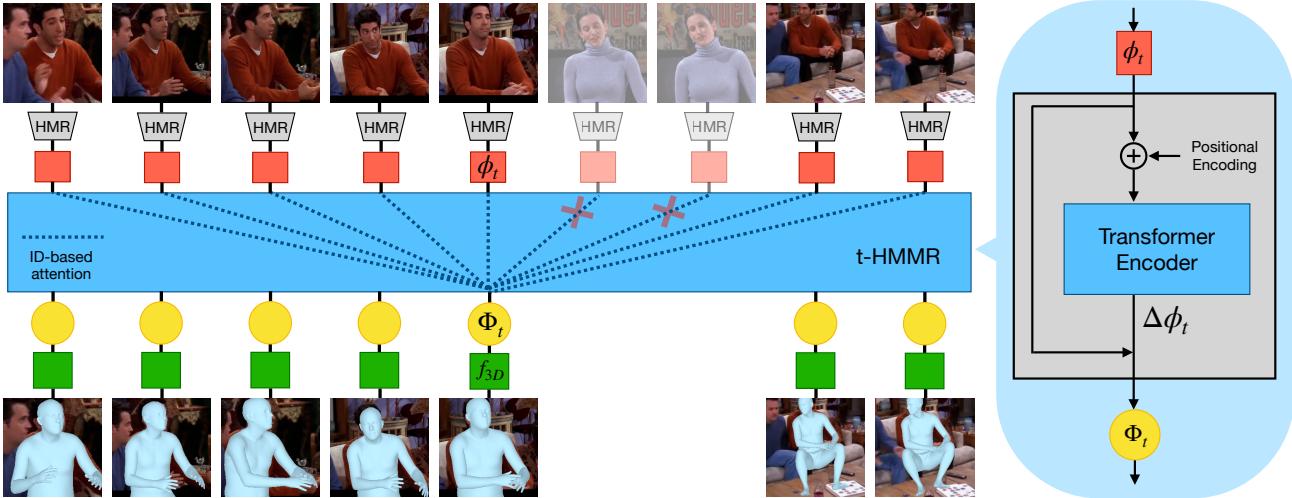


Figure 5: **Architecture of t-HMMR:** To most effectively leverage the plethora of 3D pose sequences recovered from our data, we propose t-HMMR, a human mesh and motion recovery model based on the transformer architecture. Even if the identity of interest is not present in some frames, we can still benefit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

Our transformer encoder takes as input an intermediate HMR 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 the transformer layer is a residual value $\Delta\phi_t$ added to the feature ϕ_t through a residual connection. The result of this block is the video feature representation Φ_t . This is illustrated in Figure 5.

For training the transformer encoder, following prior work [21, 22], we fix the weights of the image encoder f_{im} , and only update the temporal encoder f_{movie} and the parameter regressor f_{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_{2D}^t and L_{smp}^t respectively, for each frame t . Also, to further encourage temporal consistency, smoothness losses are applied on 3D joints $L_{\text{sm joint}}^t$ and 3D model parameters $L_{\text{sm joint}}^t$ (equivalent to equations 1 and 2 respectively).

5. Experiments

The focus of our quantitative evaluation is three-fold: we first demonstrate the quality of our multi-shot optimization results; then we investigate the effect of our data for im-

proving the performance and robustness of a single-frame human mesh recovery model; finally we examine our temporal model and validate the importance of leveraging the transformer architecture. In the next paragraph, we present in detail each one of these aspects.

5.1. Implementation details

For our HMR baseline, we retrain an HMR model on the standard datasets (Human3.6M [17], COCO [26], MPII [1]) using pseudo ground truth SMPL parameters from SPIN [23], and incorporating the cropping augmentation scheme [19, 37]. We use this baseline to initialize our multi-shot optimization and for ablative experiments. After the offline multi-shot optimization, our final HMR model is trained with the same strategy, including the data from MS-AVA. For the t-HMMR model, we freeze the image encoder of HMR, as done in [21, 22], for computational reasons, and train the temporal encoder and 3D regressor only.

5.2. Quantitative evaluation

Multi-shot optimization: Our multi-shot optimization integrates information across the shot boundary to improve pose reconstruction. To evaluate its success, we use the cross-shot PCK metric as discussed in Section 3.3. As a sanity check, we compare with two optimization based baselines, one that considers a single frame only [6], and one that considers a temporal sequence without shot changes [4, 22]. The results are presented in Figure 6, where multi-shot optimization outperforms the two baselines. As expected, the multi-shot optimization outperforms the two baselines, since the formulation of the objective function al-

lows the integration of information from multiple shots, and the large gap in performance indicates that it is very successful in this regard. Qualitative examples of this behavior are presented in Figure 7 and in the Supplementary.

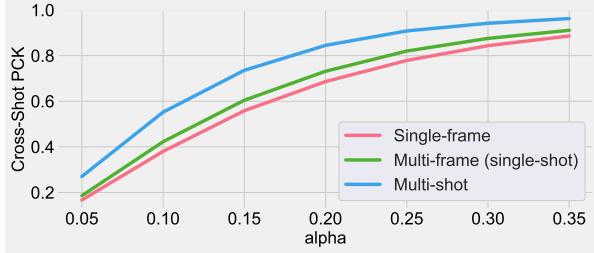


Figure 6: **Multi-shot optimization evaluation on AVA.** We show cross-shot PCK for varying thresholds α . Our multi-shot optimization outperforms optimization baseline applied on a single-frame or single-shot (multiple frames that do not span shot changes).

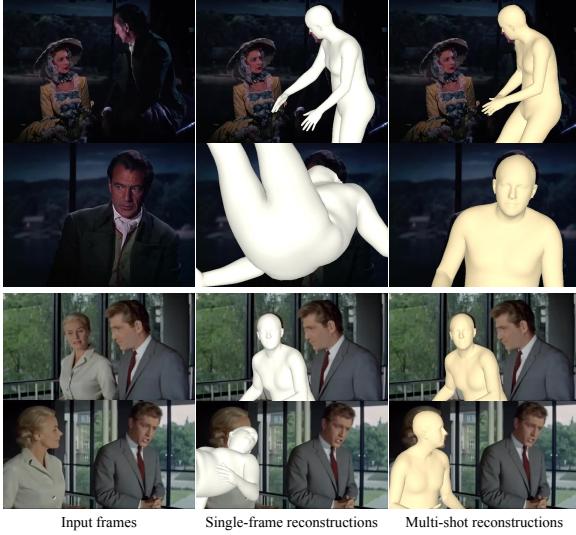


Figure 7: **Qualitative effect of our multi-shot optimization.** Although a single frame baseline fails on the more challenging frames with heavy truncation (center), our multi-shot optimization leverages information from the less ambiguous frame across the shot boundary to get a more accurate 3D reconstruction.

Single-frame model: The introduced MS-AVA dataset provides a rich source of data for training our single-frame model for human mesh recovery. Considering the nature of the data, they could help improving the robustness of our model, particularly when it comes to truncation, a common mode of failure for the state-of-the-art models [37]. We demonstrate this on the Partial Humans dataset [37] in Table 2. Our model trained with the MS-AVA data outperforms the state-of-the-art approaches. Many of these methods are trained primarily with full body images so there are significant failures under truncation, whereas our network

Method	PCKh on Partial Humans Uncropped \uparrow			
	VLOG	YouCook	Instr	Cross
HMR [20]	81.2	93.6	86.9	92.7
GraphCMR [24]	65.7	80.1	77.5	79.3
SPIN [23]	73.4	85.1	85.6	85.5
Partial Humans* [37]	68.7	95.4	77.9	91.1
HMR (retrained)	86.9	96.8	92.4	96.2
+ MS-AVA data	90.3	98.9	94.1	98.2

Table 2: **PCKh results for different single-frame models on the uncropped subset of the Partial Humans dataset [37].** Best performance is achieved when we train on MS-AVA data. Note * operates in the harder setting, which uses the entire image as an input while others operate on cropped bounding boxes.

Method	PA-MPJPE
HMR [20]	81.3
SPIN [24]	59.2
HMR (retrained)	59.2
+ MS-AVA data	57.8

Table 3: **Single-frame regression evaluation on 3DPW.** The numbers are PA-MPJPE expressed in mm. Best performance is achieved when we include our MS-AVA data in training.

can be robust in these cases too. This robustness is illustrated in Figure 8, where we provide qualitative results for different methods on the “Uncropped” subset of Partial Humans. Moreover, our HMR model trained on MS-AVA outperforms our baseline trained with synthetic cropping augmentation, underlining the effect of our MS-AVA data at achieving better robustness. Finally, we also evaluate on the 3DPW dataset [46] in Table 3, where we compare with the most relevant approaches and observe the same trends.

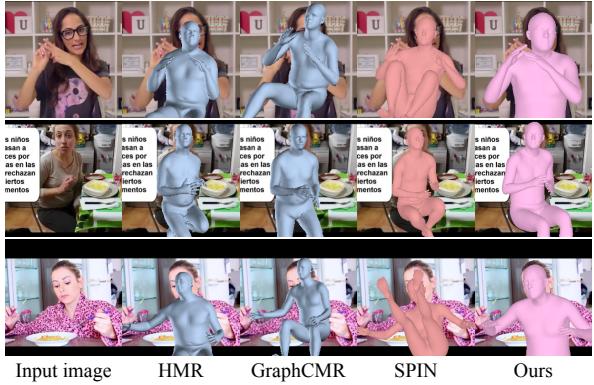


Figure 8: **Comparison against recent methods on Partial Humans dataset [37].** Qualitative comparison with recent state of the art on the Partial Humans dataset [37]. Our model is significantly more robust in the presence of truncations.

Temporal model: For our temporal model, first we investigate the transformer encoder *from an architecture viewpoint*, factoring out the effect of challenging/incomplete data. To this end, we compare it directly with other architecture choices for temporal encoders, i.e., the convolutional encoder of HMMR [21] and the recurrent encoder of VIBE [22]. For a direct comparison, we follow the exact training data/schedule with VIBE using their public implementation, and replacing only the temporal encoder. We perform multiple training runs and plot validation performance in Figure 9. If we consider the best performance across all iterations, all models achieve similar results. However, the convolutional and recurrent encoders tend to diverge after a few epochs, while the transformer encoder is more stable. Simultaneously, this version still achieves state-of-the-art results (PA-MPJPE of 56.1mm vs 56.5mm for VIBE on the 3DPW test set), even though it is trained on less data than the full VIBE model.

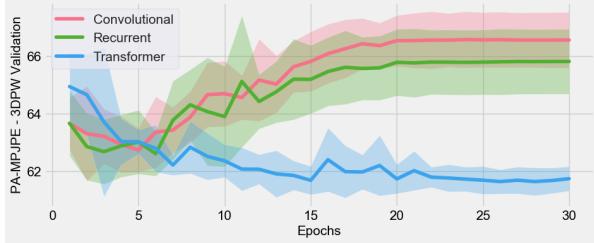


Figure 9: **Comparison of different temporal encoders on the VIBE training setting** [22]. Performance is tracked over multiple runs. Although all encoders can achieve good performance at least once during the whole training schedule, transformer training is more stable and converges consistently to a good model.

Next, we also demonstrate the suitability of the transformer model in cases of *learning and testing on incomplete data*, e.g., on movies. In these cases, the architectures other than the transformers do not have a straightforward way to handle missing frames. We experimented with different alternatives, and we found that padding missing frames with zero features performed the best for the convolutional and the recurrent encoder. Using this strategy, we report cross-shot PCK results on AVA for different encoder choices in Table 4. In the first setting, we test on the clean sequences (“test: clean” - consecutive sequences without missing frames), trained using our clean sequences (“train: clean”) from AVA. In this setting all models perform well, however, when we add the sequences with missing frames in the training (“train: all”), the transformer model benefits the most from the additional training data. In the second setting, we test on all sequences (“test: all”). The transformer still performs the best in this more challenging case.

Finally, in Figure 10 we provide some example reconstructions of our temporal t-HMMR model, in comparison

Temporal Model	train: clean	train: all	train: all
	test: clean	test: all	test: all
Convolutional	68.8	69.7	67.5
Recurrent	70.8	70.8	68.5
Transformer	70.7	73.8	70.7

Table 4: **Multi-frame evaluation on AVA, for different training/testing protocols.** The numbers are cross-shot PCK values. Transformers enjoy the largest benefit by incorporating sequences with missing data in the *training*, while they also perform the best when *tested* on sequences with missing data.

with the single frame HMR model, both trained on MS-AVA. While our HMR obtains reasonable results, output from t-HMMR is more consistent as it has a larger temporal context.

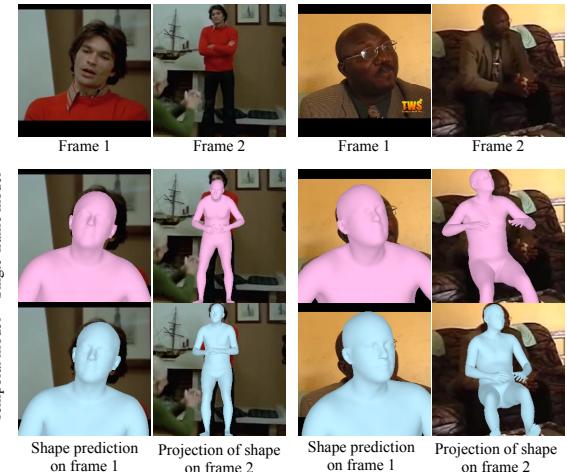


Figure 10: **Effect of t-HMMR model.** While the single-frame HMR prediction for frame 1 can be inconsistent with frame 2, our temporal t-HMMR model successfully integrates information over a temporal window and estimates a body pose for frame 1 that is consistent with frame 2.

6. Conclusion

We introduce a new task of 3D human reconstruction from multiple shots, proposing a framework to generate training data from edited media like movies, and use them in different human mesh recovery tasks. Our contributions span all three major forms of Human Mesh Recovery: off-line iterative optimization, single-view prediction, and temporal prediction. Our experiments demonstrate the significance of our contributions in each of these aspects. Future work can investigate further the available data from MS-AVA for other downstream applications or apply our data generation framework in other edited media.

Acknowledgements: This research was supported by BAIR sponsors.

References

- [1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. [6](#)
- [2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. [3](#)
- [3] Daniel Arijon. Grammar of the film language. 1976. *Hastings House*, 1976. [2](#)
- [4] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *CVPR*, 2019. [3, 4, 6](#)
- [5] Roger G Barker and Herbert F Wright. Midwest and its children: The psychological ecology of an american town. 1955. [2](#)
- [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, 6](#)
- [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, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. *arXiv preprint arXiv:2008.09047*, 2020. [3](#)
- [9] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *Excessicsa*, pages 210–227, 2020. [3](#)
- [10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. [4](#)
- [11] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Kosecka, and Ziyan Wu. Hierarchical kinematic human mesh recovery. *arXiv preprint arXiv:2003.04232*, 2020. [3](#)
- [12] 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, 3, 4](#)
- [13] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *ICCV*, 2009. [3](#)
- [14] Qingqiu Huang, Wentao Liu, and Dahua Lin. Person search in videos with one portrait through visual and temporal links. In *ECCV*, 2018. [4](#)
- [15] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. *arXiv preprint arXiv:2007.10937*, 2020. [3](#)
- [16] 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](#)
- [17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. [3, 4, 5, 6](#)
- [18] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020. [3](#)
- [19] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. [3, 6](#)
- [20] 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, 7](#)
- [21] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. [3, 5, 6, 8](#)
- [22] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. [3, 4, 5, 6, 8](#)
- [23] 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, 6, 7](#)
- [24] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. [7](#)
- [25] 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](#)
- [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*. Springer, 2014. [6](#)
- [27] 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](#)
- [28] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3D human motion estimation via motion compression and refinement. *arXiv preprint arXiv:2008.03789*, 2020. [3](#)
- [29] 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. [3, 5](#)
- [30] 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. *arXiv preprint arXiv:2008.03713*, 2020. [3](#)
- [31] 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](#)
- [32] 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
- [33] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 3
- [34] 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
- [35] 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
- [36] Anyi Rao, Lining 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
- [37] Chris Rockwell and David F Fouhey. Full-body awareness from partial observations. In *ECCV*, 2020. 3, 6, 7
- [38] 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
- [39] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatiariis, 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
- [40] Leonid Sigal, Alexandru O Balan, and Michael J Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010. 3
- [41] Ahmed AA sman, Timo Bolkart, and Michael J Black. STAR: Sparse trained articulated human body regressor. In *ECCV*, 2020. 3
- [42] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. *arXiv preprint arXiv:2008.08474*, 2020. 3
- [43] 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. 3, 5
- [44] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017. 3
- [45] 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. 5
- [46] 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. 3, 4, 5, 7
- [47] 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
- [48] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUMI: Generative 3D human shape and articulated pose models. In *CVPR*, 2020. 3
- [49] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRac: Joint 3D pose and shape estimation by dense render-and-compare. In *ICCV*, 2019. 3
- [50] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 35(12):2878–2890, 2012. 4
- [51] Andrei Zanfir, Elisabeta Marinou, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes—the importance of multiple scene constraints. In *CVPR*, 2018. 3
- [52] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3D human dynamics from video. In *ICCV*, 2019. 3
- [53] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 3, 4, 5