Full-Body Awareness from Partial Observations

Chris Rockwell[0000−0003−3510−5382] and David F. Fouhey[0000−0001−5028−5161]

University of Michigan, Ann Arbor
{cnris,fouhey}@umich.edu

Abstract. There has been great progress in human 3D mesh recovery and great interest in learning about the world from consumer video data. Unfortunately current methods for 3D human mesh recovery work rather poorly on consumer video data, since on the Internet, unusual camera viewpoints and aggressive truncations are the norm rather than a rarity. We study this problem and make a number of contributions to address it: (i) we propose a simple but highly effective self-training framework that adapts human 3D mesh recovery systems to consumer videos and demonstrate its application to two recent systems; (ii) we introduce evaluation protocols and keypoint annotations for 13K frames across four consumer video datasets for studying this task, including evaluations on out-of-image keypoints; and (iii) we show that our method substantially improves PCK and human-subject judgments compared to baselines, both on test videos from the dataset it was trained on, as well as on three other datasets without further adaptation.

Keywords: Human Pose Estimation

1 Introduction

Consider the images in Fig. 1: what are these people doing? Are they standing or sitting? While a human can readily recognize what is going on in the images, having a similar understanding is a severe challenge to current human 3D pose estimation systems. Unfortunately, in the world of Internet video, frames like these are the rule rather than rarities since consumer videos are recorded not with the goal of providing clean demonstrations of people performing poses, but are instead meant to show something interesting to people who already know how to parse 3D poses. Accordingly, while videos from consumer sharing sites may be a useful source of data for learning how the world works [2, 14, 59, 62], most consumer videos depict a confusing jumble of limbs and torsos flashing across the screen. The goal of this paper is to make sense of this jumble.

Current work in human pose estimation is usually not up to the challenge of the jumble of Internet footage. Recent work in human pose estimation [3, 9, 24, 35, 38] is typically trained and evaluated on 2D and 3D pose datasets [4, 19, 21, 30, 37] that show full human poses from level cameras often in athletic settings Fig. 2 (left). Unfortunately, Internet footage tends to be like Fig. 2 (right), and frequently only part of the body is visible to best show off how to perform a

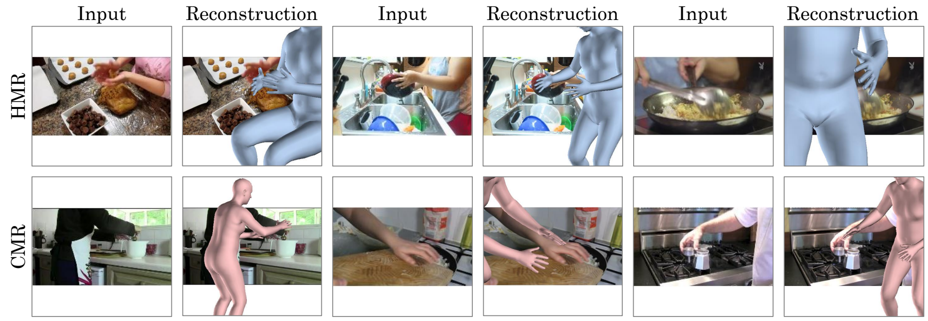

Fig. 1: We present a simple but highly effective framework for adapting human pose estimation methods to highly truncated settings that requires no additional pose annotation. We evaluate the approach on HMR [24] and CMR [26] by annotating four Internet video test sets: VLOG [14] (top-left, top-middle), Cross-Task [62] (top-right, bottom-left), YouCookII [59] (bottom-middle), and Instructions [2] (bottom-right).

task or highlight something of interest. For instance, on VLOG [14], all human joints are visible in only 4% of image frames. Meanwhile, all leg keypoints are *not* visible 63% of the time, and head keypoints such as eyes are *not* visible in about 45% of frames. Accordingly, when standard approaches are tested on this sort of data, they tend to fail catastrophically, which we show empirically.

We propose a simple but surprisingly effective approach in Section 3 that we apply to multiple forms of human mesh recovery. The key insight is to combine both cropping *and* self-training on confident video frames: cropping introduces the model to truncation, video matches context to truncations. After pre-training on a cropped version of a standard dataset, we identify reliable predictions on a large unlabeled video dataset, and promote these instances to the training set and repeat. Unlike standard self-training, we add crops, which lets confident full-body predictions (identified via [5]) provide a training signal for challenging crops. This approach requires no extra annotations and takes $<$ 30k iterations of additional training (with total time $<$ 8 hours on a single RTX2080 Ti GPU).

We demonstrate the effectiveness of our approach on two human 3D mesh recovery techniques – HMR [24] and CMR [26] – and evaluate on four consumer-video datasets – VLOG [14], Instructions [2], YouCookII [59], and Cross-Task [62]. To lay the groundwork for future work, we annotate keypoints on 13k frames across these datasets and provide a framework for evaluation in and out of images. In addition to keypoints, we evaluate using human-study experiments. Our experiments in Section 4 demonstrate the effectiveness of our method compared to off-the-shelf mesh recovery and training on crops from a standard image dataset (MPII). Our approach improves PCK both *in-image* and *out-of-image* across methods and datasets: e.g., after training on VLOG, our approach leads to a 20.7% improvement on YouCookII over off-the-shelf HMR and a 10.9% improvement over HMR trained on crops (with gains of 36.4% and 19.1% on out-of-image keypoints) Perceptual judgments by annotators show similar gains:

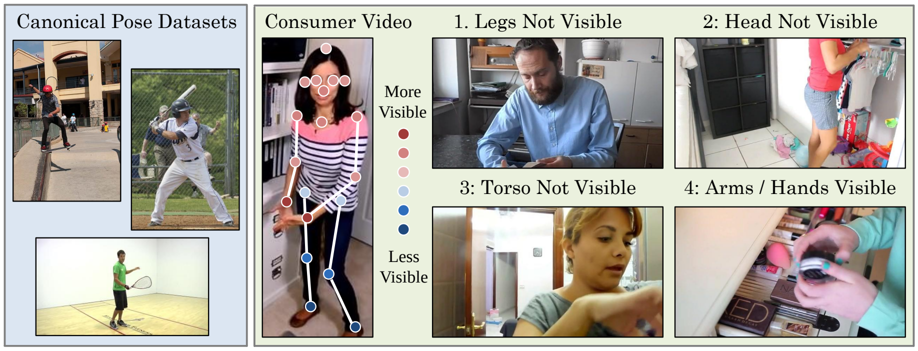

Fig. 2: Partially Visible Humans. Consumer video, seen in datasets like VLOG [14], Instructions [62], or YouCook2 [59], is considerably different from canonical human pose datasets. Most critically, only part of a person is typically visible within an image, making pose estimation challenging. In fact, all keypoints are only visible in 4% of VLOG test set images, while all leg joints are not visible 61% of the time. Four of the most common configurations of visible body parts are listed above.

e.g., on Cross-Task, our proposed method improves the chance of a CMR output being rated as correct by 25.6% compared to off-the-shelf performance.

2 Related Work

Human Pose Estimation In the Wild: Human pose estimation has improved substantially in recent years due in part to improved methods for 2D [9, 17, 38, 51, 56] and 3D [1, 28, 35, 43, 45, 60] pose, which typically utilize deep networks as opposed to classic approaches such as deformable part models [7, 10, 12, 58]. Performance of such pose models also relies critically on datasets [4, 18, 19, 21, 30, 34, 37, 48]. By utilizing annotated people *in-the-wild*, methods have moved toward understanding realistic, challenging settings, and become more robust to occlusion, setting, challenging pose, and scale variation [3, 36, 37, 40, 41, 61].

However, these *in-the-wild* datasets still rarely encounter close, varied camera angles common in consumer Internet video, which can result in people being only partially within an image. Furthermore, images that do contain truncated people are sometimes filtered out [24]. As a result, the state-of-the-art on common benchmarks performs poorly in consumer videos. In this work, we utilize the unlabeled video dataset VLOG to improve in this setting.

3D Human Mesh Estimation: A 3D mesh is a rich representation of pose, which is employed for the method presented in this paper. Compared to keypoints, a mesh represents a clear understanding of a person's body invariant to global orientation and scale. A number of recent methods [6, 24, 27, 42, 52, 55] build this mesh by learning to predict parametric human body models such as

SMPL [31] or the closely-related Adam [23]. To increase training breadth, some of these methods train on 2D keypoints [24, 42] and utilize a shape prior.

The HMR [24] model trains an adversarial prior with a variety of 2D keypoint datasets to demonstrate good performance *in-the-wild*, making it a strong candidate to extend to more challenging viewpoints. CMR [26] also produces strong results using a similar *in-the-wild* training methodology. We therefore apply our method to both models, rapidly improving performance on Internet video.

Understanding Partially-Observed People: Much of the prior work studying global understanding of partially-observed people comes from ego-centric action recognition [11, 29, 33, 47, 49]. Methods often use observations of the same human body-parts between images, typically hands [11, 29, 33], to classify global activity. In contrast, our goal is to predict pose, from varied viewpoints.

Some recent work explores ego-centric pose estimation. Recent setups use cameras mounted in a variety of clever ways, such as chest [20, 46], bike helmet [44], VR goggles [50], and hat [57]. However, these methods rely on camera always being in the same spot relative to the human to make predictions. On the other hand, our method attains global understanding of the body by training on entire people to reason about unseen joints as it encounters less visible images.

Prior work also focuses on pose estimation specifically in cases of occlusion [15, 16]. While this setting requires inference of non-visible joints, it does not face the same scale variation occurring in consumer video, which can contain people much larger than the image. Some recent work directly addresses truncation. Vosoughi and Amer predict truncated 3D keypoints on random crops of Human3.6M [54]. In concurrence with our work, Exemplar Fine-Tuning [22] uses upper-body cropping to improve performance in Internet video [34]. Nevertheless, consumer Internet video (Fig. 2) faces more extreme truncation. We show cropping alone is not sufficient for this setting; rather cropping *and* self-training on confident video frames provides the best results.

3 Approach

Our goal is the ability to reconstruct a full human-body 3D mesh from an image of part or all of a person in consumer video data. We demonstrate how to do this using a simple but effective self-training approach that we apply to two 3D human mesh recovery models, HMR [24] and CMR [26]. Both systems can predict a mesh by regressing SMPL [31] parameters from which a human mesh can be generated, but work poorly on this consumer video data.

Our method, shown in Fig. 3, adapts each method to this challenging setting of partial visibility by sequentially self-training on confident mesh and keypoint predictions. Starting with a model trained on crops from a labeled dataset, the system makes predictions on video data. We then identify confident predictions using the equivariance technique of Bahat and Shakhnarovich [5]. Finally, using the confident examples as pseudo-ground-truth, the model is trained to map crops of the confident images to the full-body inferences, and the process of

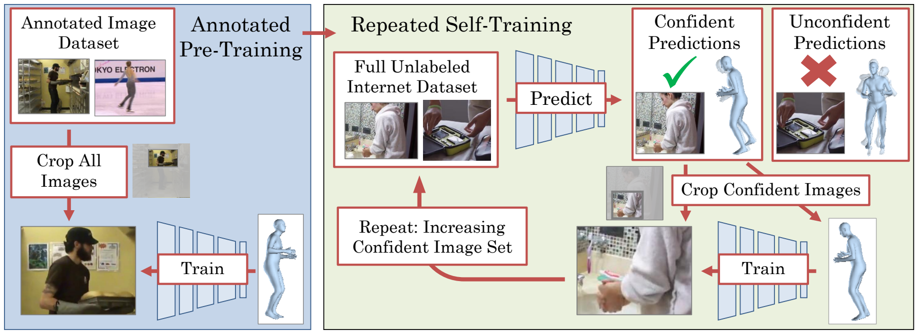

Fig. 3: Our method adapts human pose models to truncated settings by self-training on cropped images. After pre-training using an annotated pose dataset, the method applies small translations to an unlabeled video dataset and selects predictions with consistent pose predictions across translations as pseudo-ground-truth. Repeating the process increases the training set to include more truncated people.

identifying confident images and folding them into a training set is continued. Our only assumption is that we can identify frames containing a single person (needed for training HMR/CMR). In this paper, we annotate this for simplicity but assume this can be automated via an off-the-shelf detection system.

3.1 Base Models

Our base models [24, 26] use SMPL [31], which is a differentiable, generative model of human 3D meshes. SMPL maps parameters Θ to output a triangulated mesh $\mathbf{Y}$ with $N = 6980$ vertices. Θ consists of parameters $[\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{R}, \mathbf{t}, \mathbf{s}]$: joint rotations $\boldsymbol{\theta} \in \mathbb{R}^{69}$, shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, global rotation $\mathbf{R}$, global translation $\mathbf{t}$, and global scale $\mathbf{s}$. We abstract each base model as a function f mapping an image I to a SMPL parameter Θ. As described in [24], the SMPL parameters can be used to yield a set of 2D projected keypoints $\mathbf{x}$.

Our training process closely builds off the original methods: we minimize a sum of losses on a combination of projected 2D keypoints $\hat{\mathbf{x}}$, predicted vertices $\hat{\mathbf{Y}}$, and SMPL parameters $\hat{\Theta}$. The most important distinctions are, we assume we have access to SMPL parameters Θ for each image, and we train on all annotated keypoints, even if they are outside the image. We describe salient differences between the models and original training below.

HMR [24]: Kanazawa *et al.* use MoSh [32, 53] for their ground truth SMPL loss. However, this data is not available in most images, and thus the model relies primarily on keypoint loss. Instead, we train directly on predicted SMPL rotations, available in all images; we find L_1 loss works best. To encourage our network to adapt to poses of new datasets, we do not use a discriminator loss. We also supervise $(\mathbf{t}, \mathbf{s})$, though experiments indicated this did not impact performance (less than 1% difference on keypoint results). The loss for a single datapoint is:

$$L = \left\| [\boldsymbol{\theta}, \mathbf{R}, \boldsymbol{\beta}, \mathbf{t}, \mathbf{s}] - [\hat{\boldsymbol{\theta}}, \hat{\mathbf{R}}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{t}}, \hat{\mathbf{s}}] \right\|_1 + \|\mathbf{x} - \hat{\mathbf{x}}\|_1 \tag{1}$$

CMR [26]: CMR additionally regresses predicted mesh, and has intermediate losses after the Graph CNN. We do not change their loss, other than by always using 3D supervision and out-of-image keypoints. It is distinct from our HMR loss as it uses L_2 loss on keypoints and SMPL parameters, and converts $\boldsymbol{\theta}$ and $\mathbf{R}$ to rotation matrices for training [39], although they note conversion does not change quantitative results. The loss for a single datapoint is:

$$L = \left\| [\boldsymbol{\theta}, \mathbf{R}] - [\hat{\boldsymbol{\theta}}, \hat{\mathbf{R}}] \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right\|_2^2 + \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|_1 \tag{2}$$

such that $\lambda = 0.1$, each norm is reduced by its number of elements, and keypoint and mesh losses are also applied after the Graph CNN. While Kolotouros *et al.* train the Graph CNN before the MLP, we find the pretrained model trains well with both losses simultaneously.

3.2 Iterative Adaptation to Partial Visibility

Our approach follows a standard self-training approach to semi-supervised learning. In self-training, one begins with an **initial model** $f_0 : \mathcal{X} \rightarrow \mathcal{Y}$ as well as a collection of unlabeled data $U = \{u : u \in \mathcal{X}\}$. Here, the inputs are images, outputs SMPL parameters, and model either CMR or HMR. The key idea is to use the inferences of each round's model f_i to produce labeled data for training the next round's model f_{i+1}. More specifically, at each iteration t, the model f_t is applied to each element of U, and a **confident prediction subset** $C \subseteq U$ is identified. Then, predictions of model f on elements are treated as new ground-truth for training the next round model f_{i+1}. In standard self-training, the new training set is the original unlabeled inputs and model outputs, or $\{(c, f_i(c)) : c \in C\}$. In our case, this would never learn to handle more cropped people, and the training set is thus augmented with **transformations** of the confident samples, or $\{(t(c), t(f_i(c))) : c \in C, t \in T\}$ for some set of crops T. The new model f_{i+1} is **retrained** and the process is repeated until convergence. We now describe more concretely what we mean by each bolded point.

Initial Model: We begin by training the pretrained HMR and CMR models on MPII (Fig. 3, left) such that we apply cropping transformations to images and keypoints. SMPL predictions from full images are used for supervision, and are typically very accurate considering past training on this set. This training scheme is the same as that used for self-training (Fig. 3, right), except we use MPII ground truth keypoints instead of pseudo-ground truths.

Identifying Confident Predictions: In order to apply self-training, we need to be able to find confident outputs of each of our SMPL-regressing models. Unfortunately, it is difficult to extract confidence from regression models because there is no natural and automatically-produced confidence measure unlike in classification where measures like entropy provide a starting point.

We therefore turn to an empirical result of Bahat and Shakhnarovich [5] that invariance to image transformations is often indicative of confidence in neural networks. Put simply, confident predictions of networks tend to be more invariant to small transformations (e.g., a shift) than non-confident predictions. We apply this technique in our setting by examining changes of parameters after applying small translational jitter: we apply the model f to copies of the image with the center jittered 10 and 20 pixels and look at joint rotation parameters $\boldsymbol{\theta}$. We compute the variance of each joint rotation parameter across the jittered samples, then average the variances across joints. For HMR, we define confident samples as ones with a variance below 0.005 (chosen empirically). For CMR, for simplicity, we ensure that we have the same acceptance rate as HMR of 12%; this results in a similar variance threshold of 0.004.

Applying Transformations: The set of inputs and confident pseudo-label outputs that can be used for self-training is not enough. We therefore apply a family of crops that mimic empirical frequencies found in consumer video data. Specifically, crops consist of 23% most of body visible, 29% legs not visible, 10% head not visible, and 22% only hands or arms visible. Examples of these categories are shown in Fig. 2. Although proportions were chosen empirically from VLOG, other consumer Internet video datasets considered [2, 59, 62] exhibit similar visibility patterns, and we empirically show that our results generalize.

Retraining: Finally, given the set of samples of crops of confident images and corresponding full bodies, we retrain each model.

3.3 Implementation Details

Our cropping procedure and architecture is detailed in supplemental for both HMR [24] and CMR [26]; we initialize both with weights pretrained on *in-the-wild* data. On MPII, we continue using the same learning rate and optimizer used by each model (1e-5 for HMR, 3e-4 for CMR, both use Adam [25]) until validation loss converges. Training converges within 20k iterations in both cases.

Next, we identify confident predictions as detailed above on VLOG. We use the subset of the hand-contact state dataset containing single humans, which consists of 132k frames in the train + validation set. We note we could have used a simple classifier to filter by visible people in a totally unlabeled setting. Our resulting confident train + validation set is 15k images. We perform the same cropping transformations as in MPII, and continue training with the same parameters. Validation loss converges within 10k iterations. We repeat this semi-supervised component one additional time, and the new train + validation set is of approximately size 40k. Training again takes less than 10k iterations.

4 Experiments

We now describe a set of experiments done to investigate the following experimental questions: (1) how well do current 3D human mesh recovery systems work on consumer internet video? (2) can we improve the performance of these

Fig. 4: Randomly sampled **positive** and **negative** predictions by number of keypoints visible, as identified by workers. Images with fewer keypoints visible are typically more difficult, and our method improves most significantly in these cases (Table 2).

systems, both on an absolute basis and in comparison to alternate simple models that do not self-train? We answer these questions by evaluating the performance of a variety of adaptation methods applied to both HMR [24] and CMR [26] on four independent datasets of Internet videos. After introducing the datasets (Sec. 4.1) and experimental setup (Sec. 4.2), we describe experiments on VLOG (Sec. 4.3), which we use for our self-training video adaptation. To test the generality of our conclusions, we then repeat the same experiments on three other consumer datasets without any retraining (Sec. 4.4). We validate our choice of confidence against two other methods in (Sec. 4.5).

4.1 Datasets and Annotations

We rely on four datasets of consumer video from the Internet for evaluating our method: VLOG [14], Instructions [2], YouCookII [59], and Cross-Task [62]. Evaluation on VLOG takes place on a random 5k image subset of the test set detailed in Sec. 3.3. For evaluation on Instructions, YouCookII, and Cross-Task, we randomly sample test-set frames (Instructions we sample from the entire dataset, which is used for cross-validation), which are filtered via crowd-workers by whether there is a single person, and then randomly subsample 5k subset.

Table 1: Joint visibility statistics across the four consumer video datasets that we use. Across multiple consumer video datasets, fully visible people are exceptionally rare ($< 3\%$), in contrast to configurations like an upper torso or only pieces of someone's arms, or much of a body but no head. Surprisingly, the most likely to be visible joint is actually wrists, more than 2x more likely than hips and 5x more likely than knees.

	Independent Joint Statistics							Joint Joint Statistics			
	Ankle	Knee	Hip	Wrist	Elbow	Neck + Should.	Head + Face	Fully Visible	Upper Torso	Only Arms	All But Head
Average	7.0	12.9	31.9	71.9	50.8	53.9	51.1	2.8	26.2	31.8	18.8
VLOG [14]	10.5	20.0	34.0	71.5	54.5	61.0	53.3	4.0	32.0	24.0	14.0
Instructions [2]	14.0	24.5	32.5	73.5	52.0	44.7	43.3	5.0	14.0	32.0	18.0
YouCook II [59]	0.0	1.0	30.0	71.0	45.5	53.7	52.7	0.0	28.0	39.0	24.0
Cross-Task [62]	3.5	6.0	31.0	71.5	51.0	56.3	55.0	2.0	31.0	32.0	19.0

Finally, to enable automatic metrics like PCK, we obtain joint annotations on all four datasets. We annotate keypoints for the 19 joints reprojected from HMR, or the 17 COCO keypoints along with the neck and head top from MPII. Annotations are crowd-gathered by workers who must pass a qualification test and are monitored by sentinels, and is detailed in supplemental. We show statistics of these joints in Table 1, which show quantitatively the lack of visible keypoints. In stark contrast to canonical pose datasets, the head is often not visible. Instead, the most frequently visible joints are wrists.

4.2 Experimental Setup

We evaluate our approaches as well as a set of baselines that test concrete hypotheses using two styles of metrics: 2D keypoint metrics, specifically PCK measured on both in-image joints as well as via out-of-image joints (via evaluation on crops); and 3D Mesh Human Judgments, where crowd workers evaluate outputs of the systems on an absolute or relative basis.

2D Keypoint Metrics: Our first four metrics compare predicted keypoints with annotated ones. Our base metric is PCK @ 0.5 [4], the percent of keypoints within a threshold of 0.5 times head segment, the most commonly reported threshold on MPII. Our first metric, *Uncropped PCK*, is performance on images where the head is visible to define PCK. We choose PCK since head segment length is typically undistorted in our data, as opposed to alternates where identifying a stable threshold is difficult: PCP [13] is affected by our high variance in body 3D orientation, and PCPm [4] by high inter-image scale variation.

PCK is defined only on images where the head is visible (a shortcoming we address with human judgment experiments). In addition to being a subset, these frames are not representative of typical visibility patterns in consumer video (as shown in Fig. 2 and Table 1), so we evaluate on crops. We sample crops to closely match the joint visibility statistics of each entire annotated test set (detailed in supplemental). We can then evaluate *In-Image PCK*, or PCK on joints in the cropped image. Because the original image contains precise annotations of joints not visible in the crop, we can also evaluate *Out-of-Image PCK*, or PCK on joints outside the crop. *Total PCK* is PCK on both. We calculate PCK on each image and then average over images. Not doing this gives significantly more weight to images with many keypoints in them, and ignores images with few.

3D Mesh Human Judgments: While useful, keypoint metrics like PCK suffer from a number of shortcomings. They can only be evaluated on a subset of images: this ranges from 37% of images from Instructions to 50% of images in Cross-Task. Moreover, these subsets are not necessarily representative, as argued before. Finally, in the case of out-of-image keypoints, PCK does not distinguish between plausible predictions that happen to be incorrect according to the fairly exacting PCK metric, and implausible guesses. We therefore turn to human judgments, measuring results in absolute and comparative terms.

Mesh Score/Absolute Judgment: We show workers an image and single mesh, and ask them to classify it as largely correct or not (precise definition in supplemental), from which we can calculate Percentage of Good Meshes: the proportion of predicted meshes workers consider good. Predictions from all methods are aggregated and randomly ordered, so are evaluated by the same pool of workers.

Relative Judgment: As a verification, we also perform A/B testing on HMR predictions. We follow a standard A/B paradigm and show human workers an image and two meshes in random order and ask which matches the image better with the option of a tie; when workers cannot agree, we report this as a tie.

Baselines: We compare our proposed model with two baselines to answer a few scientific questions.

Base Method: We compare with the base method being used, either HMR [24] or CMR [26], without any further training past their original pose dataset training sets. This both quantifies how well 3D pose estimation methods work on consumer footage and identifies when our approach improves over this model.

Crops: We also compare with a model trained on MPII Crops (including losses on out-of-image keypoints). This tests whether simply training the model on crops is sufficient compared to also self-training on Internet video.

4.3 Results on VLOG

Our first experiments are on VLOG [14], the dataset that we train on. We begin by showing qualitative results, comparing our method with a number of baselines in Fig. 5. While effective on full-body cases, the initial methods perform poorly on truncated people. Training on MPII Crops prepares the model to better identify truncated people, but self-training on Internet video provides the model context clues it can associate with people largely outside of images — some of the largest improvements occur when key indicators such as sinks and tables (Fig. 5) are present. In Fig. 6, the model identifies distinct leg and head poses outside of images given minute difference in visible pose and appearance.

Human 3D Mesh Judgments: We then consider human 3D Mesh Judgments, which quantitatively confirm the trends observed in the qualitative results. We report the frequency that each method's predictions were rated as largely correct on the test set, broken down by the number of visible joints, in Table 2. Our

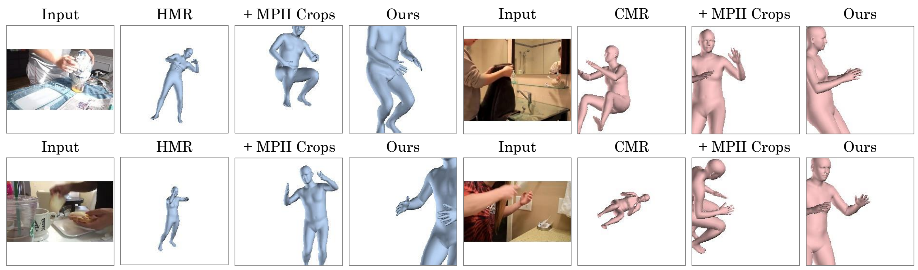

Fig. 5: Selected comparison of results on VLOG [14]. We demonstrate sequential improvement between ablations on HMR (left) and CMR (right). Training on MPII Crops prepares the model for truncation, while self-training provides context clues it can associate with full-body pose, leading to better predictions, particularly outside images.

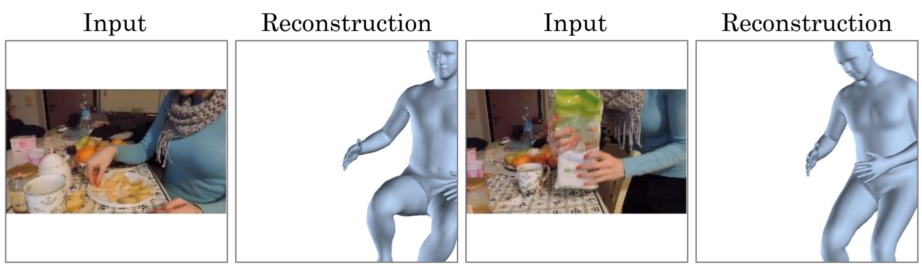

Fig. 6: Shots focused on hands occur often in consumer video. While the visible body may look similar across instances, full-body pose can vary widely, meaning keypoint detection is not sufficient for full-body reasoning. After self-training, our method learns to differentiate activity such as standing and sitting given similar visible body.

approach *always* outperforms using the base method, and only is outperformed by Crops on full or near-full keypoint visibility. These performance gains are particularly strong in the less-visible cases compared to both the base method and crops. For instance, by using our technique, HMR's performance in highly truncated configurations (1-3 Keypoints Visible) is improved by 23.7 points compared to the base and 11.0 compared to using crops.

2D Keypoints: We next evaluate keypoints, reporting results for all four variants in Table 3. On cropped evaluations that match the actual distribution of consumer video, our approach produces substantial improvement, increasing performance overall for both HMR and CMR. On the *uncropped images* where the head of the person is visible (which is closer to distributions seen on e.g., MPII), our approach remains approximately the same for HMR and actually improves by 8.6% for CMR. We note our method underperforms within cropped images on HMR. There are two reasons for this: first, supervising on out-of-image keypoints encourages predictions outside of images, sacrificing marginal in-image performance gains. Second, the cost of supervising on self-generated keypoints

Table 2: Percentage of Good Meshes on VLOG, as judged by human workers. We report results on *All* images and examine results by number of visible keypoints.

	HMR [24]							CMR [26]						
	By # of Visible Joints							By # of Visible Joints						
	1-3	4-6	7-9	10-12	13-15	16-19	All	1-3	4-6	7-9	10-12	13-15	16-19	All
Base	19.2	52.7	70.1	80.1	85.2	82.1	60.6	13.6	37.9	53.4	68.0	79.0	74.8	51.1
Crops	31.9	68.7	76.9	86.8	91.0	**85.9**	69.4	33.7	65.8	75.7	82.5	88.4	**80.9**	67.5
Full	**42.9**	**72.1**	**82.8**	**89.6**	**92.3**	83.1	**73.9**	**40.9**	**71.2**	**80.2**	**86.0**	**89.2**	79.5	**71.2**

Table 3: PCK @ 0.5 on VLOG. We compute PCK on the 1.8k image VLOG test set, in which the head is fully visible, as *Uncr. Total*. These images are then *Cropped* to emulate the keypoint visibility statistics of the entire dataset, on which we can calculate PCK *In* and *Out* of cropped images, and their union *Total*.

Method	HMR [24]				CMR [26]			
	Cropped			Uncr.	Cropped			Uncr.
	Total	In	Out	Total	Total	In	Out	Total
Base	48.6	65.2	14.7	68.5	36.1	50.2	13.2	49.5
Crops	51.6	**65.3**	24.2	**68.8**	47.3	58.1	26.2	**59.5**
Ours	**55.9**	61.6	**38.9**	68.7	**50.9**	**60.3**	**34.6**	58.1

is reduced precision in familiar settings. Nevertheless, CMR improves enough using semi-supervision to still increase on in-image-cropped keypoints.

4.4 Generalization Evaluations

We now test generalization to other datasets. Specifically, we take the approaches evaluated in the previous section and apply them directly to Instructions [2], YouCookII [59], and Cross-Task [62] *with no further training*. This tests whether the additional learning is simply overfitting to VLOG. We show qualitative results of our system applied to these datasets in Fig. 7. Although the base models also work poorly on these consumer videos, simply training on VLOG is sufficient to produce more reasonable outputs.

3D Mesh Judgments: This is substantiated quantitatively across the full dataset since, as shown in Table 4, HMR and CMR perform poorly out-of-the-box. Our approach, however, can systematically improve their performance without any additional pose annotations: gains over the best baseline range from 4.5 percentage points (CMR tested on Instructions) to 10.7 percentage points (HMR tested on YouCookII). Our outputs are systematically preferred by humans in A/B tests (Table 5): our approach is 4.6x – 8.9x more likely to be picked as preferable compared to the base system than the reverse, and similarly 2.4x – 7.8x more likely to be picked as preferable to crops than the reverse.

2D Keypoints: Finally, we evaluate PCK. Our approach produces strong performance gains on two out of the three datasets (YouCookII and Cross-Task),

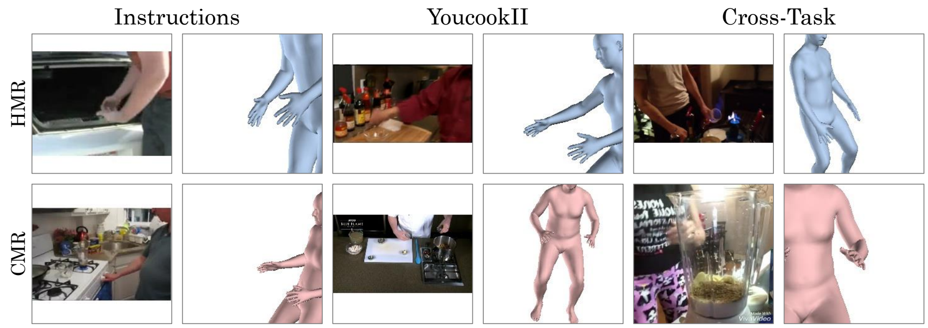

Fig. 7: Results on External Datasets. While our method trains on Internet Vlogs, performance generalizes to other Internet video consisting of a variety of activities and styles; specifically instructional videos and cooking videos.

Table 4: Percentage of Good Meshes on External Datasets, as judged by human workers. We report results on *All* images and in the case of few visible keypoints.

	Instructions [2]				YouCook II [59]				Cross-Task [62]			
	HMR		CMR		HMR		CMR		HMR		CMR	
	1-6	All	1-6	All	1-6	All	1-6	All	1-6	All	1-6	All
Base	10.7	42.2	7.4	30.9	15.9	54.6	8.5	41.8	13.6	52.6	7.7	37.9
Crops	25.5	53.8	28.8	52.5	24.9	60.8	24.3	60.2	22.3	59.0	21.5	57.7
Full	**37.3**	**60.5**	**35.2**	**57.0**	**43.4**	**71.5**	**39.9**	**68.5**	**37.0**	**68.1**	**31.0**	**63.5**

while its performance is more mixed on Instructions *relative to MPII Crops*. We hypothesize the relatively impressive performance of MPII Crops is due to 40% of this dataset consisting of car mechanical fixes. These videos frequently feature people bending down, for instance while replacing car tires. Similar activities such as swimming are more common in MPII than VLOG. The corresponding array of outdoor scenes also provides less context to accurately infer out-of-image body parts. Yet, strong human judgment results (Table 4, 5) indicate training on VLOG improves coarse prediction quality, even in this setting.

4.5 Additional Comparisons

To validate our choice of confidence, we consider two alternative criteria for selecting confident images: agreement between HMR and CMR SMPL parameters, and agreement between HMR and Openpose [8] keypoints. For fair comparison, implementations closely match our confidence method; full details and tables are in supplemental. Compared to both, our system does about the same or better across datasets, but does not require running two systems. Agreement with CMR yields cropped keypoint accuracy of 1.5-2.7% lower, and uncropped accuracy of 0.6% higher - 0.6% lower. Agreement with Openpose is stronger on uncropped images: 0.3%-2.4% higher, but weaker on uncropped: 1.3%-3.5% lower.

Table 5: A/B testing on All Datasets, using HMR. For each entry we report how frequently (%) the row wins/ties/loses to the column. For example, row 2, column 6 shows that our full method is preferred 47% of the time over a method trained on MPII Crops, and MPII Crops is preferred over the full method just 6% of the time

Method	VLOG [14]		Instructions [2]		YouCookII [59]		Cross-Task [62]	
	Base	Crops	Base	Crops	Base	Crops	Base	Crops
Crops	53/28/19	-	56/28/16	-	49/35/16	-	46/39/15	
Full	63/23/15	45/43/12	65/21/14	40/43/17	62/32/7	47/47/6	57/36/7	41/53/7

Table 6: PCK @ 0.5 on External Datasets. We compute PCK in test set images in which the head is fully visible. These images are then cropped to emulate the keypoint visibility statistics of the entire dataset, on which we can calculate PCK on predictions outside the image.

Method	Instructions [2]				YouCookII [59]				Cross-Task [62]			
	HMR [24]		CMR [26]		HMR [24]		CMR [26]		HMR [24]		CMR [26]	
	Total	Out	Total	Out	Total	Out	Total	Out	Total	Out	Total	Out
Base	42.0	19.6	32.8	17.1	56.0	27.7	44.0	26.9	56.1	20.3	44.1	19.8
MPII Crops	**50.6**	33.7	**47.9**	**33.9**	65.8	45.0	65.0	48.6	62.9	32.5	61.9	38.2
Ours	48.7	**36.4**	44.8	33.7	**76.7**	**64.1**	**70.7**	**58.5**	**74.5**	**57.2**	**66.9**	**47.9**

We additionally consider performance of our model to the model after only the first iteration of VLOG training, through A/B testing (full table in supplemental). In all four datasets, the final method is 1.7x – 2.8x more likely to be picked as preferable to the model after only one round than the reverse.

5 Discussion

We presented a simple but effective approach for adapting 3D mesh recovery models to the challenging world of Internet videos. In the process, we showed that current methods appear to work poorly on Internet videos, presenting a new opportunity. Interestingly, while CMR outperforms HMR on Human3.6M, the opposite is true on this new data, suggesting that performance gains on standard pose estimation datasets do not always translate into performance gains on Internet videos. Thanks to the new annotations across the four video datasets, however, we can quantify this. These keypoint metrics are validated as a measure for prediction quality given general agreement with human judgement metrics in extensive testing. We see getting systems to work on consumer videos, including both the visible and out-of-image parts, as an interesting and impactful challenge and believe our simple method provides a strong baseline for work in this area.
Acknowledgments: This work was supported by the DARPA Machine Common Sense Program. We thank Dimitri Zhukov, Jean-Baptiste Alayrac, and Luowei Zhou, for allowing sharing of frames from their datasets, and Angjoo Kanazawa and Nikos Kolotouros for polished and easily extended code. Thanks to the members of Fouhey AI Lab and Karan Desai for the great suggestions!

References

1. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: CVPR (2015)
2. Alayrac, J.B., Bojanowski, P., Agrawal, N., Laptev, I., Sivic, J., Lacoste-Julien, S.: Unsupervised learning from narrated instruction videos. In: CVPR (2016)
3. Alp Güler, R., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: CVPR (2018)
4. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
5. Bahat, Y., Shakhnarovich, G.: Confidence from invariance to image transformations. arXiv preprint arXiv:1804.00657 (2018)
6. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV (2016)
7. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV (2009)
8. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018)
9. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR (2017)
10. Desai, C., Ramanan, D.: Detecting actions, poses, and objects with relational phraselets. In: ECCV (2012)
11. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: ICCV. pp. 407–414 (2011)
12. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
13. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR (2008)
14. Fouhey, D.F., Kuo, W., Efros, A.A., Malik, J.: From lifestyle VLOGs to everyday interactions. In: CVPR (2018)
15. Ghiasi, G., Yang, Y., Ramanan, D., Fowlkes, C.C.: Parsing occluded people. In: CVPR (2014)
16. Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., Fei-Fei, L.: Towards viewpoint invariant 3D human pose estimation. In: ECCV (2016)
17. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: ECCV (2016)
18. Ionescu, C., Li, F., Sminchisescu, C.: Latent structured models for human pose estimation. In: ICCV (2011)
19. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. TPAMI (2013)
20. Jiang, H., Grauman, K.: Seeing invisible poses: Estimating 3D body pose from egocentric video. In: CVPR (2017)
21. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BVMC (2010)
22. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. arXiv preprint arXiv:2004.03686 (2020)

23. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3D deformation model for tracking faces, hands, and bodies. In: CVPR (2018)
24. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014)
26. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019)
27. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3D and 2D human representations. In: CVPR (2017)
28. Lee, H.J., Chen, Z.: Determination of 3D human body postures from a single view. Computer Vision, Graphics, and Image Processing **30**(2), 148–168 (1985)
29. Li, Y., Ye, Z., Rehg, J.M.: Delving into egocentric actions. In: CVPR (2015)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
31. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6) (2015)
32. Loper, M., Mahmood, N., Black, M.J.: Mosh: Motion and shape capture from sparse markers. ACM Transactions on Graphics (TOG) **33**(6), 220 (2014)
33. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: CVPR (2016)
34. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 601–617 (2018)
35. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: CVPR (2017)
36. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 International Conference on 3D Vision (3DV). pp. 506–516. IEEE (2017)
37. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3D human pose estimation with a single RGB camera. ACM Transactions on Graphics (TOG) **36**(4), 44 (2017)
38. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016)
39. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 international conference on 3D vision (3DV). pp. 484–494. IEEE (2018)
40. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: CVPR (2017)
41. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: CVPR (2017)
42. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3D human pose and shape from a single color image. In: CVPR (2018)
43. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3D human pose from 2D image landmarks. In: ECCV (2012)
44. Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.P., Schiele, B., Theobalt, C.: Egocap: egocentric marker-less motion capture with two fisheye cameras. ACM Transactions on Graphics (TOG) **35**(6), 1–11 (2016)

45. Rogez, G., Schmid, C.: Mocap-guided data augmentation for 3d pose estimation in the wild. In: Advances in neural information processing systems (2016)
46. Rogez, G., Supancic, J.S., Ramanan, D.: First-person pose recognition using egocentric workspaces. In: CVPR (2015)
47. Ryoo, M.S., Matthies, L.: First-person activity recognition: What are they doing to me? In: CVPR (2013)
48. Sigal, L., Balan, A.O., Black, M.J.: 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)
49. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV (2016)
50. Tome, D., Peluse, P., Agapito, L., Badino, H.: xR-EgoPose: Egocentric 3D human pose from an HMD camera. In: ICCV (2019)
51. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: CVPR (2014)
52. Tung, H.Y., Tung, H.W., Yumer, E., Fragkiadaki, K.: Self-supervised learning of motion capture. In: NeurIPS (2017)
53. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: CVPR (2017)
54. Vosoughi, S., Amer, M.A.: Deep 3D human pose estimation under partial body presence. In: ICIP (2018)
55. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: CVPR (2019)
56. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: ECCV (2018)
57. Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H.P., Theobalt, C.: Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. IEEE transactions on visualization and computer graphics **25**(5), 2093–2101 (2019)
58. Yang, Y., Ramanan, D.: Articulated pose estimation using flexible mixtures of parts. In: CVPR (2011)
59. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: AAAI (2018)
60. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3D human pose estimation from monocular video. In: CVPR (2016)
61. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3D human pose estimation in the wild: a weakly-supervised approach. In: ICCV (2017)
62. Zhukov, D., Alayrac, J.B., Cinbis, R.G., Fouhey, D., Laptev, I., Sivic, J.: Cross-task weakly supervised learning from instructional videos. In: CVPR (2019)