Full-Body Awareness from Partial Observations Supplemental Material

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

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

Table of Contents

1 Method

1.1 HMR Network Architecture

Our HMR network architecture is the same as the HMR model of Kanazawa *et al.* It consists of a ResNet-50 feature extractor, followed by a 3D regression refinement network, consisting of 3 fully-connected layers mapping to SMPL, global rotation, and camera parameters Θ. The fully-connected layers concatenate image features, mean SMPL parameters $\boldsymbol{\Theta}$, and default global rotation $\mathbf{R}$ and camera parameters $[\mathbf{t}, \mathbf{s}]$. FC layers also use residual links and dropout. More details can be found in the original paper. Also like HMR, we use same-padding for image inputs, although for illustrative purposes images in the paper are shown with white or black padding.

1.2 CMR Network Architecture

Our CMR network architecture is the same as the CMR model of Kolotouros *et al.* It consists first of a ResNet-50 encoder, with the final fully-connected layer removed. This outputs a 2048-D feature vector, which is attached to 3D coordinates of template mesh vertices. A series of graph convolutions then map to a single 3D mesh vertex set, and to camera parameters $[\mathbf{t}, \mathbf{s}]$. Finally, a multi-layer perceptron maps these vertices to SMPL parameters $\boldsymbol{\Theta}$ and global rotations

$\mathbf{R}$. Final predictions use camera parameters from graph convolutions $[\mathbf{t}, \mathbf{s}]$, and SMPL parameters and global rotations $[\boldsymbol{\Theta}, \mathbf{R}]$ from the MLP. More details can be found in the original paper.

1.3 Confident Predictions

As detailed in the paper, predictions are considered confident if average variance of joint rotation parameters across jittered images is less than 0.005 for HMR, chosen empirically. For simplicity, threshold for CMR is chosen so that approximately the same number of confident images are chosen, resulting in a threshold of 0.004. The five images from which predictions are averaged are:

1. the original image
2. the image translated 10 pixels to the top left, padded on the bottom right
3. the image translated 20 pixels to the top left, padded on the bottom right
4. the image translated 10 pixels to the bottom right, padded on the top left
5. the image translated 20 pixels to the bottom right, padded on the top left

1.4 Cropping

During training, the proposed method crops training and validation images into five categories: above hip, above shoulders, from knee to shoulder, around only an arm, and around only a hand. We show examples of cropping in Fig. 1. These crops correspond to common crops occurring in consumer video, displayed in Fig. 2 of the paper. Above hip corresponds to "Legs Not Visible", knee to shoulder corresponds to "Head Not Visible", and above shoulders corresponds to "Torso Not Visible". For brevity, in Fig. 2, we condensed only an arm or only a hand into one image: "Arms / Hands Visible".

During training, we sample crops with approximately the same frequency as they occur in the VLOG validation set. Proportions are: above hip in 29% of images, knee to shoulder in 10% of images, above shoulders in 16% of images, around one arm only in 9% of images, around one hand in 13% of images, and 23% of images we leave uncropped. On both MPII and VLOG, for both models, we crop to our target crops using keypoints. Ground truth keypoints are used on MPII, and reprojected keypoints from confident models are used on VLOG. Above hip crops are made from the lower of the hip keypoints. Knee to shoulder crops use the higher of knee keypoints to the bottom of shoulder keypoints. Above shoulder crops use the lower of shoulder and neck as the bottom of the crop. Elbow and wrist keypoints are used to approximate one arm and one hand crops. If keypoints used for cropping are outside of images, for simplicity we presume the image is already cropped and do not crop further. If a prospective crop would be smaller than 30 pixels, we also do not crop to prevent training on very low-resolution examples.

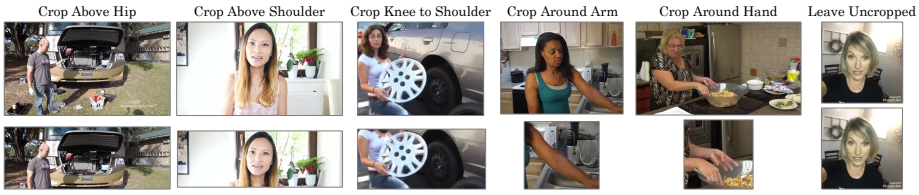


Fig. 1: Generating Cropped Sets. Training and cropped testing use keypoints to crop images to target visibility specifications. Examples of each crop specification we use are pictured. Some images are left uncropped, and sometimes predefined crops do not further crop images (right).

2 Datasets

2.1 Dataset Annotation

As detailed in the paper, for each of VLOG, Instructions, YoucookII, and Cross-Task; we subsample 5k random frames containing exactly one person. Next, we use human annotators to label human keypoints on all of these frames. The full test sets consist of images in which at least one keypoint is annotated, on which workers can agree (details in Keypoint Annotation). They are of size 4.1k, 2.2k, 3.3k, and 3.9k, for VLOG, Instructions, YoucookII, and Cross-Task, correspondingly. Instructions is notably the smallest because it has a higher proportion of images containing no human keypoints. The most common instance of this occurring is when some of a hand is visible, which does not contain a keypoint, but a wrist is not visible, which corresponds to a keypoint.

All human annotations were gathered using thehive.ai, a website similar to Amazon Mechanical Turk. Annotators were given instructions and examples of each category. Then, they were given a test to identify whether they understood the task. Workers that passed the test were allowed to annotate images. However, as they annotated, they were shown, at random, gold standard sentinel examples (i.e., that we labeled), that tested their accuracy. The platform automates the entire process. Because some workers spoke only Spanish, we put directions in both English and Spanish. English annotation instructions are provided in Subsections 5.1, 5.2, and 5.3.

Keypoint Annotation

Although traditionally disagreement on labeling ground truth is handled by thehive.ai, the company does not currently support labeling keypoints in the instance there are an ambiguous number of keypoints visible, which occurs here. For instance, consider the case where a person's elbow is at the edge of an image. Some annotators may label the elbow while others do not. To deal with annotations containing different number of visible joints, we combine predictions between workers on a given image ourselves, using the median number of joints annotated between workers, and average the locations.

Specifically, we have each image labeled three separate times. If two or three of the occurrences of the image see no joints labeled, we consider it as having no

joints visible. In other cases, we take the median number of joints visible between the three instances, and average these joints across instances when possible. If any given joint has predictions differing by a large margin (10% of image size), we do not annotate it. If all three workers disagree on number of joints visible, we consider it ambiguous and do not add it to our test set.

3D Mesh Annotation

Absolute Judgment: Workers annotated whether the mesh was largely correct or not. The percent of times they annotated correct gives Percentage of Good Meshes. For mesh scoring, all ablations were scored at the same time, to avoid possible bias between scores between models. Additionally, model outputs and images were put in random order to avoid one person seeing many outputs from the same model in a row.

Relative Judgment: For A/B testing, workers were presented with an image and two outputs. They selected the output that best matched the image. Order in which predictions were seen in relation to the image was randomized, and which outputs were compared was also presented in random order.

2.2 Dataset Details and Statistics

As explained in the paper, we evaluate keypoint accuracy on images in which the head is visible in order to calculate PCK. This results in test subsets of size 1.8k, 0.8k, 1.5k, and 1.9k, for VLOG, Instructions, YoucookII, and Cross-Task, correspondingly. These sets are not representative of the full test-set visibility statistics, and do not allow for out-of-image keypoint evaluation. Therefore, we use cropping of body-parts to closely match aggregate test set statistics. We use the same canonical crops as during training, displayed in Fig. 1. However, we explicitly choose crop proportions to closely match full test sets.

Uncropped keypoint test sets are biased since their images always contain head keypoints; needed for computing PCK. Therefore, we must crop aggressively to match full test set statistics. Furthermore, above hip and above shoulder crops are not useful to this end, as they include head keypoints. Knee to shoulder keypoints also are not optimal as they exclude leg keypoints too often, while continuing to sometimes include shoulder and neck keypoints. Instead, to match full test set statistics, we utilize crops around hands and arms frequently, while leaving some images uncropped. Statistics on full test sets, uncropped keypoint test sets, and cropped keypoint test sets are detailed in Table 1.

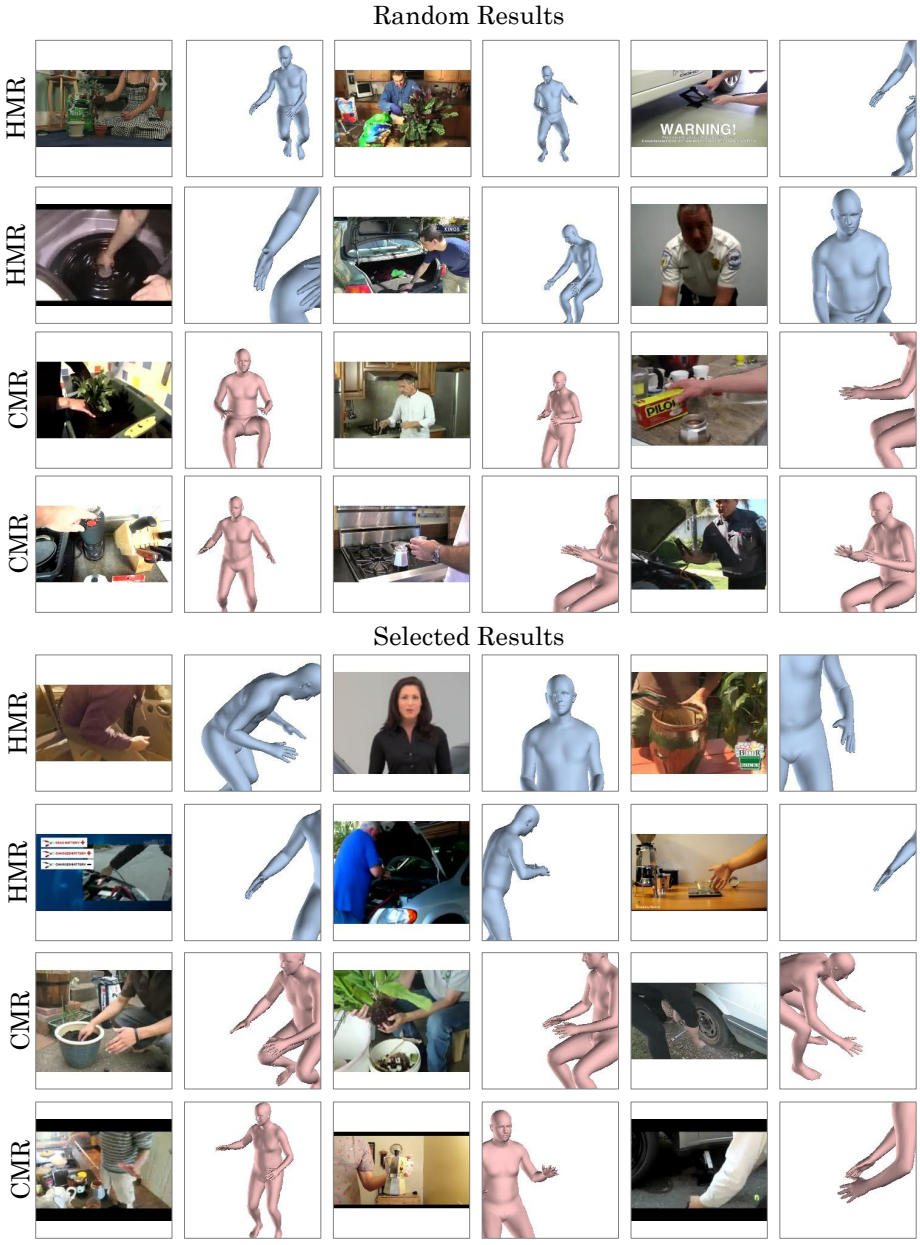
Table 1: Proportion of Visible Joints in Test Sets. Proportion of dataset images containing a particular joint for each of: Uncropped Keypoint Test Set (Uncr.), Cropped Keypoint Test Set (CR.), and Full Test Set (Full). Also, mean number of keypoints (Keypoints) per image.

	VLOG			Instructions			YoucookII			Cross-Task		
	Uncr.	Cr.	Full	Uncr.	Cr.	Full	Uncr.	Cr.	Full	Uncr.	Cr.	Full
R Ank	0.16	0.07	0.11	0.22	0.11	0.13	0.00	0.00	0.00	0.04	0.02	0.03
R Kne	0.29	0.16	0.20	0.32	0.23	0.24	0.01	0.00	0.01	0.07	0.04	0.06
R Hip	0.48	0.30	0.34	0.64	0.40	0.32	0.51	0.33	0.30	0.46	0.30	0.31
L Hip	0.48	0.31	0.34	0.66	0.43	0.33	0.51	0.35	0.30	0.46	0.30	0.31
L Kne	0.29	0.16	0.20	0.33	0.23	0.25	0.01	0.00	0.01	0.07	0.04	0.06
L Ank	0.15	0.07	0.10	0.23	0.11	0.15	0.00	0.00	0.00	0.04	0.02	0.04
R Wri	0.78	0.65	0.73	0.83	0.69	0.76	0.78	0.61	0.72	0.76	0.59	0.72
R Elb	0.75	0.48	0.55	0.87	0.57	0.52	0.79	0.50	0.45	0.78	0.50	0.51
R Sho	0.95	0.63	0.61	0.95	0.43	0.45	0.99	0.60	0.53	0.97	0.60	0.56
L Sho	0.95	0.62	0.61	0.94	0.43	0.44	0.99	0.58	0.54	0.97	0.60	0.56
L Elb	0.76	0.49	0.54	0.88	0.56	0.52	0.78	0.51	0.46	0.78	0.51	0.51
L Wri	0.76	0.66	0.70	0.83	0.70	0.71	0.76	0.61	0.70	0.75	0.61	0.71
Neck	1.00	0.67	0.61	1.00	0.46	0.45	1.00	0.62	0.54	1.00	0.63	0.57
Head Top	1.00	0.55	0.47	1.00	0.28	0.38	1.00	0.36	0.48	1.00	0.47	0.51
Nose	0.95	0.61	0.57	0.98	0.42	0.47	1.00	0.52	0.54	0.99	0.56	0.57
L Eye	0.95	0.58	0.55	0.98	0.38	0.45	1.00	0.47	0.54	0.99	0.54	0.56
R Eye	0.95	0.58	0.55	0.98	0.36	0.45	1.00	0.48	0.54	0.99	0.53	0.56
L Ear	0.93	0.56	0.53	0.96	0.35	0.43	0.99	0.47	0.53	0.99	0.53	0.55
R Ear	0.93	0.56	0.53	0.94	0.33	0.42	0.99	0.47	0.53	0.98	0.53	0.55
Keypoints	13.5	8.7	8.8	14.5	7.5	7.9	13.1	7.5	7.7	13.1	7.9	8.2

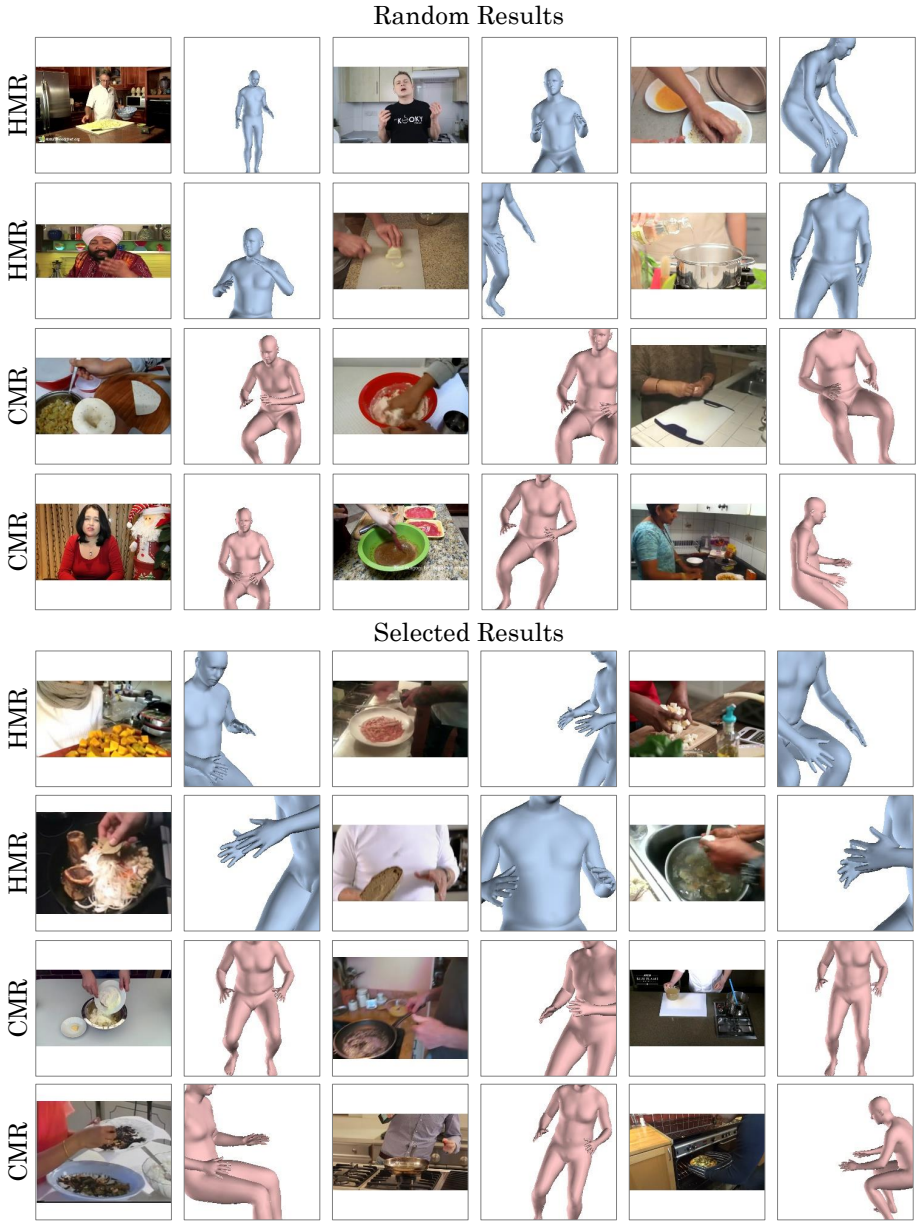
3 Additional Qualitative Results

3.1 Additional Results on VLOG

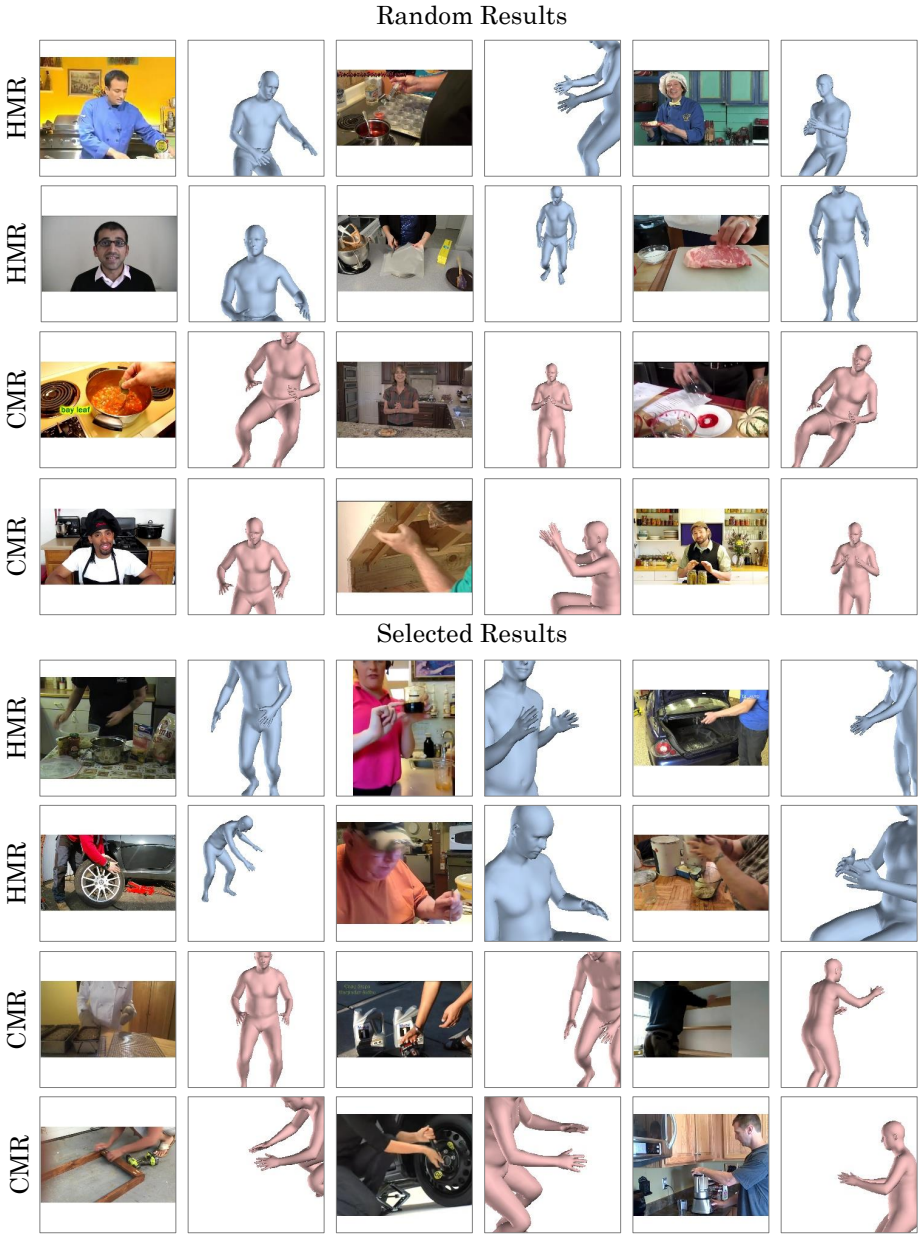
3.2 Additional Results on Instructions



3.3 Additional Results on YoucookII



3.4 Additional Results on Cross-Task



4 Additional Results

Comparison to One Round: The paper presents comprehensive comparison between our full method, the method after only MPII crops, and original baselines. As our method performs two iterations of training on VLOG, we additionally compare final performance to the method after only a single round of VLOG training in Table 2. As reported in the paper, workers prefer the model after a second iteration of training.

Table 2: A/B testing on All Datasets, using HMR. For each entry we report how frequently (%) our full method wins/ties/loses to the method trained with only one round of VLOG training. For example, column 1 shows Full is preferred to One Round 18% of the time, and One Round is preferred just 8% of the time.

Method	One Round			
	VLOG	Instructions	YouCookII	Cross-Task
Full	18/74/8	15/76/9	18/75/7	15/80/5

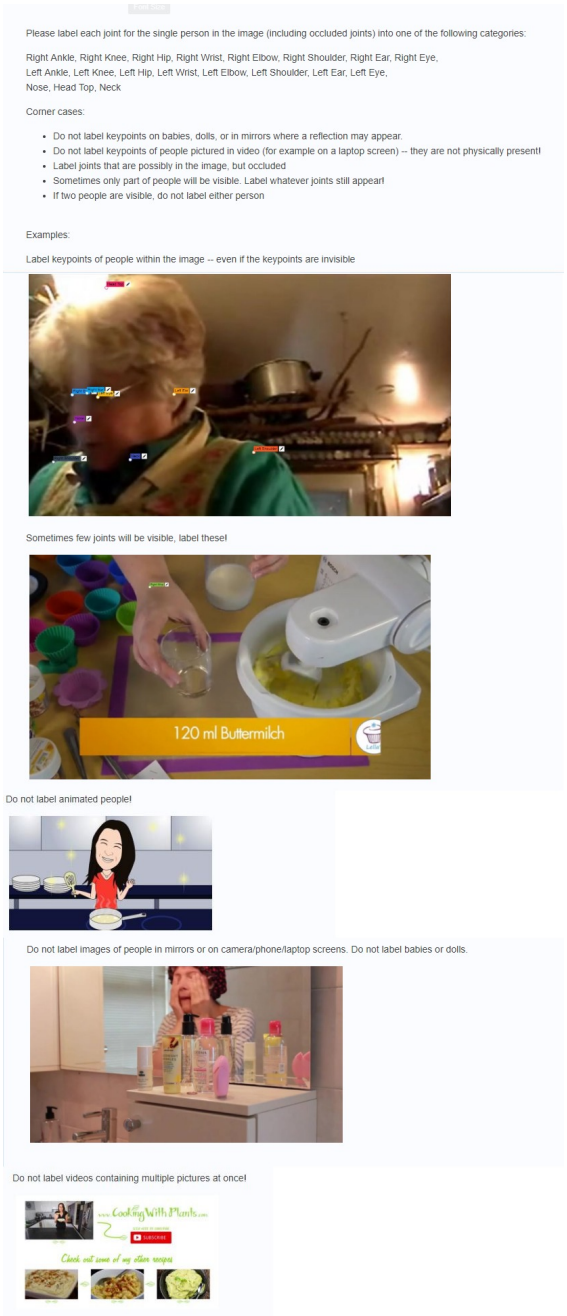
Comparison to other confidence methods: As reported in the paper, our selection of confidence also performs similarly to slightly better to agreement between CMR and HMR SMPL joints, and to agreement between HMR keypoints and Openpose keypoints. Full results are available in Table 3. CMR / HMR agreement is done in the same manner as our method, and thresholds are set to select approximately the same number of images as our method. Training takes approximately the same time, and two rounds of self-training are used. The same is true of comparison with Openpose, with the distinction Openpose can only predict keypoints inside an image, so we only consider joints both networks predict as in-image. Additionally, Openpose filters out unconfident keypoints, so we only compare joints predicted by both networks. We observe Openpose struggles especially if the face is truncated, so agreement is mostly in highly-visible settings. This leads to better uncropped keypoint accuracy, but worse cropped.

Table 3: PCK @ 0.5 on All Datasets, using HMR. 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. We also compute PCK on the uncropped images.

Method	VLOG			Instructions			YouCookII			Cross-Task		
	Cropped		Uncr.	Cropped		Uncr.	Cropped		Uncr.	Cropped		Uncr.
	Total	Out	Total	Total	Out	Total	Total	Out	Total	Total	Out	Total
CMR agreement	54.3	35.9	68.1	47.2	33.9	78.5	74.0	59.5	94.9	72.2	51.8	90.7
Openpose agreement	54.6	34.6	**71.1**	46.1	31.8	**79.8**	73.2	58.8	**95.7**	71.3	50.0	**92.2**
Ours	**55.9**	**38.9**	68.7	**48.7**	**36.4**	77.9	**76.7**	**64.1**	95.4	**74.5**	**57.2**	91.1

5 Annotation Instructions

5.1 Keypoint Annotation Instructions



5.2 Mesh Scoring Instructions

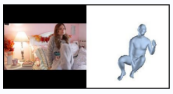
We have a system that is trying to predict human mesh corresponding to an image. We'd like to identify cases where it has a general idea of where the person is, and of their pose. Providing a perfect mesh can often be difficult, so medium errors are acceptable. Please label each mesh as good or bad.

1. Bad: Mesh does not fit image well.
   - Either the mesh seems like it could have been generated from an arbitrary image, or errors are huge.
2. Good: Mesh fits image well.
   - Errors are not huge.

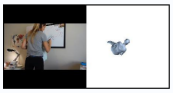
Below are some examples.

[Good] The mesh matches the below person pretty well.

[Good] This mesh isn't perfect but is fairly good.



[Bad] The below mesh has incorrect orientation, scale, location, and pose. The mesh generated essentially seems random considering the image. Therefore, it is bad.



[Bad] Here, the scale and location of the person is totally off. The mesh generated essentially seems random considering the image. Therefore, it is bad.

[Good] Larger errors can be tolerated on when a person is outside the image. Even if it isn't perfect it is clear the model has some idea of the person!

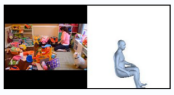
[Good] This mesh is rotated wrong, but errors are not huge.

[Good] Scale of this person should be bigger but it clearly understands the person's general orientation and pose.

[Good] Again, some errors are okay, if mesh gets general idea.

[Bad] The model doesn't seem to be getting the idea of this image. Erros in scale, location and orientation are large.

[Bad] Errors are pretty large here: the legs are wrong and direction is wrong, and location is wrong.



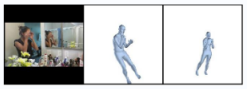
5.3 A/B Testing Instructions

Please choose which mesh best fits the person in image.

Ties are allowed if multiple predictions are essentially the same. If one is even slightly better, it should be selected.

Below are some examples.

Correct: **left**. The left mesh more closely fits scale and location of person in the image.

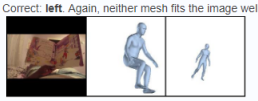


Correct: **tie**. Both mesh are extremely similar and fit the image equally well.

Correct: **right**. The right mesh has better scale and location of the person.

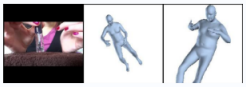
Correct: **left**. Although neither mesh fits the image well, the scale, location and pose of the left image is closer to the actual person.

Correct: **left**. Again, neither mesh fits the image well, but the scale of the left image is closer to that in the image.



Correct: **right**. Although not within the image, it is likely this girl is sitting. Therefore, the mesh on the right is better.

Correct: **right.** Here, the image on the right has better scale.



Correct: **left**. Although both predictions are similar, the one on the right has an arm going through the body which is less realistic than the arms on the left.