**Abstract**

* HMR 2.0 – state of the art performance of tracking and reconstructing humans on video via transformers
* Converts monocular video of 2D into 3D reconstructions (‘meshes’) to improve tracking during occlusion

**Introduction**

* Process is essentially obtaining 3D meshes with unprecedented accuracy from single images, then bridging these reconstructions across frames for videos.
* An advancement of recent work on human mesh tracking (HMR -- <https://arxiv.org/pdf/1712.06584.pdf> and 3D tracking via PHALP -- <https://arxiv.org/pdf/2112.04477.pdf>) via transformerization
  + Transformerization – converting CNNs and LSTMs (Long Short-Term Memory networks)
* Relies on creating a tracking entity for each person in the video and then tracking them over time via HMR over single frames
* Contributions:
  + Propose a ‘transformerized’ architecture for HMR (HMR 2.0) that outperforms existing approaches for 3D body pose estimation while not using domain-specific designs
  + Design 4DHumans – achieving state-of-the-art results for human tracking
  + Improved pose accuracy leads to improved action recognition

**Related Work**

* Human Mesh Recovery from a Single Image
  + Relies on the original HMR model
  + There are many advancements on HMR relying on domain-specific decisions to increase performance, but HMR 2.0 uses no domain-specific knowledge
* Human Mesh and Motion Recovery from Video
  + Most designs simply have a temporal encoder to fuse features across frames
    - VIBE, MEVA, and TCMR to name a few
  + Limitation of this – often work best when tracking is simple, with few occlusions
  + 4DHumans solves the occlusion problem by tracking in 3D
* Tracking People in Video
  + Current state-of-the-art performance maps video into 3D human reconstructions via HMR (ie PHALP), then tracks across 3D representations over time
  + PHALP used in HMR 2.0 and 4DHumans
* Action Recognition
  + Recent successes in the area use appearance features from raw video input (SlowFast and MViT) or extract features from body pose information (Po-Tion and JMRN)
  + More recently, the top performer has fused video-based features with features from 3D human pose estimates
    - This is the pipeline used for action recognition in this paper

**Reconstructing People**

* Body Model
  + Uses an SMPL (<https://files.is.tue.mpg.de/black/papers/SMPL2015.pdf>) model to reconstruct the human body
* Camera
  + Perspective camera model is used with fixed focal length and intrinsic parameter matrix
  + Rotation matrix and translation vector used to map points in the SMPL (human body model) space to the image
    - Since the SMPL model already is rotated as it is a human, the model assumes the rotation matrix is the identity matrix and only try to predict camera translation
* Human Mesh Reconstruction
  + Goal is to create a 3D reconstruction of a person by predicting their 3D pose and shape parameters
* Architecture
  + “end-to-end transformer architecture with no domain-specific design choices” that outperforms all existing architectures, even those with domain-specific and complex design choices
  + **ViT – Vision Transformer**
    - Transformer modified to operate on an image
    - Steps, summarized:
      * Converts an input image into input tokens
      * Input tokens processed via the transformer to get output tokens
      * Output tokens processed via the transformer decoder
  + **Transformer decoder**
    - Use a standard transformer decoder with multi-head self-attention (read more about this in the well-known paper <https://arxiv.org/abs/1706.03762>)
    - Processes input token by cross-attending to output image tokens
    - Provides a linear readout of the model parameters
* Losses
  + Typical best practice in HMR literature, model is trained via combination of 2D losses, 3D losses, and a discriminator

1. Standard MSE loss when ground-truth (annotated) SMPL parameters are known
2. Additionally:
   1. If accurate 3D keypoint annotations available, use an L1 loss of the 3D keypoints
   2. If accurate 2D keypoint annotations available, use an L1 loss of the predicted 3D keypoints using known camera transformations of the 2D keypoints
   3. Discriminator is trained for each factor of the body model to ensure the pose predicted is valid for a human

* Working with Unlabeled Datasets
  + Use an off-the-shelf detector and body keypoint estimator to get corresponding 2D points
  + Fit an SMPL mesh to the 2D keypoints to get ‘pseudo-ground truth’ SMPL parameters

**Tracking People**

* Build upon PHALP, which tracks people across time from features of their 3D reconstructions
  + Create 3D reconstructions
  + Create tracklets of people by fusing across frames
    - Create predictions for each person in the next frame, then select the best match between these predictions and the reconstructions in the next frame
  + Because you’re working in 3D space, account for occlusions is simplified and effective
* Train the pose prediction model via masking (SHOW THE IMAGE) of random pose tokens
  + Allows to make predictions AND bridge gaps due to occlusions (amodal completion)

**Experiments**

* Key Results:
  + HMR 2.0 outperforms previous methods of 2D and 3D pose accuracy
  + 4DHumans achieves state-of-the-art performance
  + Accurate poses lead to improved accuracy of action recognition tasks
* Setup
  + Used typical datasets for training (Human3.6M, MPI-INF-3DHP, etc) from previous work
* Pose Accuracy
  + HMR 2.0a is a reduced model training only on the typical datasets and for less time than HMR 2.0b
  + 3D Metrics
    - HMR 2.0a can outperform all previous baselines across all metrics
  + 2D Metrics
    - HMR 2.0b had improved performance over all, including HMR 2.0a, especially when testing on more unusual poses
* Tracking
  + Evaluation Metrics
    - Evaluated based on ID Switches, Multiple Object Tracking Accuracy (MOTA), the ratio of correct detections over the average number of ground-truth detections (IDF1), and Higher Order Tracking Accuracy (HOTA)
  + HMR 2.0 is in the top group of models for 3D pose estimation ( SHOW RESULTS TABLE )
  + 4DHumans improves across all evaluated metrics for tracking ( SHOW RESULTS TABLE )
* Action Recognition
  + Trained a separate action classifier using SMPL poses as inputs
  + HMR 2.0 SMPL poses outperformed baselines across all different metrics

**Conclusion**

* Proposed HMR 2.0, an end-to-end transformerized version of existing work for Human Mesh Recovery
* Proposed 4DHumans for state-of-the-art results in tracking
* Applied HMR 2.0 to action recognition and it improved on existing baseline scores
  + As well, by making the PHALP more generalizable, HMR 2.0 can work with a wider range of input videos/images
* Suggested future limitations to address:
  + Use of SMPL creates limitations; improved models would allow for modeling hand pose and facial expressions, or a greater age variation
  + Current model considers people independently, so contact between people is not always captured well
  + Lower input resolution reduces the quality of the 3D reconstructions