Self-supervised human depth estimation from monocular videos

Feitong Tan[1]
Ping Tan[1]
Hao Zhu[2]
Zhaopeng Cui[3]
Marc Pollefeys[3]
Siyu Zhu[4]

[1]Simon Fraser University
[2]Nanjing University
[3]ETH Zurich
[4]Alibaba AI Labs

[Link to paper]

Aim

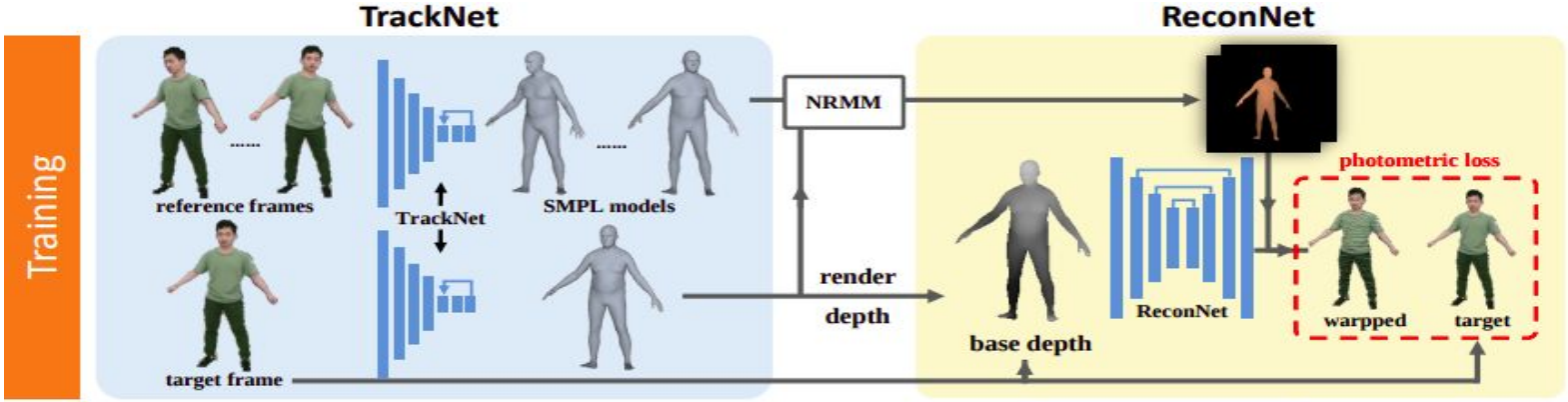
A self-supervised approach to monocular human depth reconstruction from RGB videos by training on YouTube videos without known depth or camera parameters

Approach

Minimizing photo-consistency loss, between a reference frame and its neighboring frames after estimating warping 3D non-rigid motion of the human body.

Method

1. Estimate an SMPL model separately for each video frame. From this estimate base depth map for each frame.
2. Compute the non-rigid body motion between SMPL model of reference frame and models of neighboring frames.
3. Warp the reference frame to neighboring frames using this estimated motion and minimize a photo-consistency loss. Use this loss to train a network that learns a residual detail map which is added to base depth map to obtain final depth map.



Step 1: TrackNet

- Estimate an SMPL model at each video frame to capture the human pose and rough shape.
- Fine-tuned a pretrained HMR model by capturing a set of videos with ‘ground-truth’ SMPL coefficients generated by DoubleFusion. TrackNet can use any method that outputs an SMPL model i.e. not limited to accuracy of HMR. The following loss is used for fine-tuning

$$L_{tn} = L_{para} + \theta_p L_{J\_pos} + \theta_r L_{J\_rot}$$

- Output of TrackNet: 85-D vector, with 82 parameters as SMPL coefficients and 3 parameters for the weak-perspective camera model(R, t, s).

Step 1: TrackNet

- Most state of the art methods perform weak perspective projection:

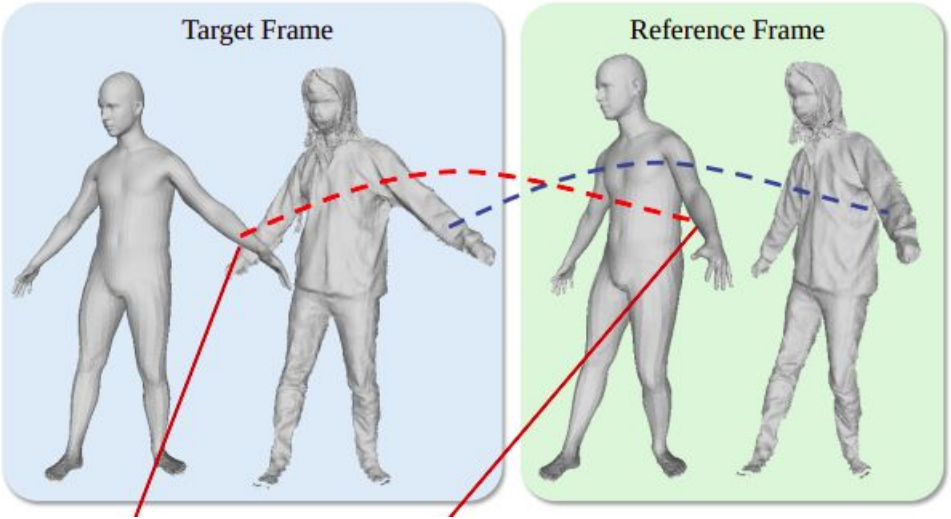
$$\hat{\mathbf{x}} = s\Pi(RX(\boldsymbol{\theta}, \boldsymbol{\beta})) + t$$

- However, general perspective projection is more accurate and better for photometric loss
- Assume a medium focal length and convert translation vector as follows:

$$V_{tras.} = [t_x, t_y, \frac{f_c}{\frac{1}{2} * img\_size * s}],$$

Step 2: Non-Rigid Motion Model

- Estimate 3D dense spatial transformation of SMPL vertices between neighboring frames. Final output: 2D Motion map similar to scene flow
- Assumption: Non-rigid transformation between the SMPL models is the same as that between the detailed shapes

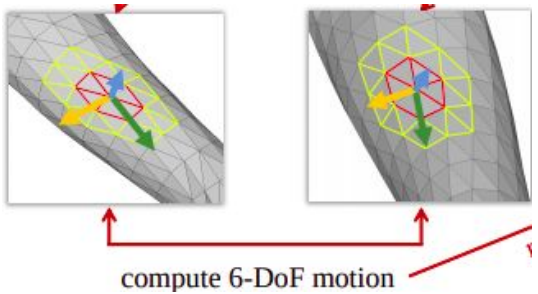


Per-vertex 6-DoF transformation

- Since the SMPL models at two neighboring frames share the same topology, they have explicit per-vertex correspondence.
- The per-vertex transformation can be computed by registering *n* two-ring neighboring vertices in the target and reference model. The rotation matrix, *R* and translation vector, *t* can be computed by:

$$R = \arg\min \sum_{i=1}^{n} ||R(v_t^i - v_t^c) - (v_r^i - v_r^c)||^2,$$

$$t = v_r - R * v_t,$$



- This is computed for every vertex to create a 3D motion field. A 2D motion map is generated by ray tracing over this motion field. A depth map of the SMPL model is also rendered called as base depth, D_t.

Effect of baseline length and occlusion for photo-consistency

- A baseline length is defined for each pixel in the motion map, which is the magnitude of its translation *t*. The mean baseline over all pixels is computed for each target-reference frame pair and remove pairs with mean baseline < 0.5m.
- To handle occlusions due to large motion, a validation mask M_r is computed, where a pixel is ‘valid’ if it is visible in both target and reference view and has a baseline length larger than 5 cm.

Step 3: ReconNet

- ReconNet computes the detail depth layer i.e an additive depth layer to capture texture details is added to the base depth to estimate the final body shape and pose.
- Input: 512×512 RGB image concat with zero-median rendered base depth map from the estimated SMPL model.
- Output: 256×256 depth offset map (residual depth map). The depth offset is added to base depth to get final depth map.
- A variant of U-Net using residual blocks in encoder and decoder with skip connections was used. The encoder has 6 down-sampling layers, while the decoder has 5 up-sampling layers. A final sigmoid layer is applied at the end to regularize the output from -10 to 10 cm.

Losses: Photo-consistency

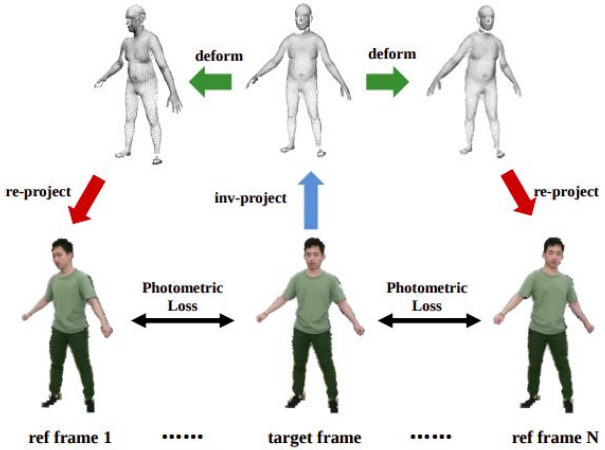


Figure 3: We first inverse-project the composed depth to point clouds, then deform them with non-rigid motion map and finally reproject deformed point clouds back to reference image for bilinear sampling.

- After composing the residual and base depth to obtain the full depth map, a photo-consistency loss is calculated.
- The inverse-warped images from each reference frame can be synthesized according to:

$$p_r \sim K T_{t\to r}(p_t) D(p_t) K^{-1} p_t$$

- Loss is computed as:

$$L^r_{photo} = \left(\alpha \frac{1 - SSIM_{cs}(I_t, \hat{I}^r_t)}{2} + (1-\alpha) \parallel I_t - \hat{I}^r_t \} \parallel \right) \otimes M_r$$

$$SSIM_{cs} = \frac{\sigma_{xy} + c}{\sigma_x^2 + \sigma_y^2 + c}$$

Losses

- Photo-consistency loss is summed over all reference images for better robustness

$$L_{photo} = \sum_{r=1, r \neq t}^{N} L^{r}_{photo}.$$

- Since a human shape is being estimated, the gradient of the final depth map must be close to the base shape, which leads to the following smoothness term:

$$L_{smooth} = \sum_{p_t} |\nabla D_{detail}(p_t) - \nabla D_{base}(p_t)|.$$

- The final depth must also be similar to base depth, which leads to the following regularization term:

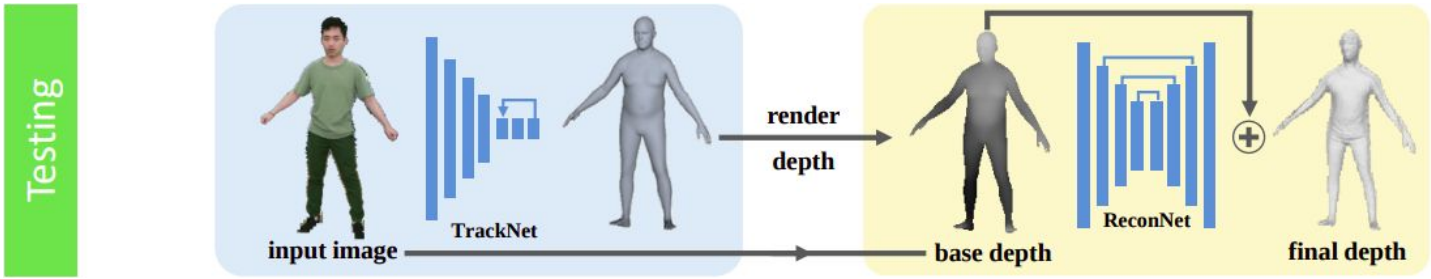
$$L_{regularizer} = \sum_{p_t} |D_{detail}(p_t) - D_{base}(p_t)|.$$

- The final learning objective function is:

$$L = L_{photo} + \gamma_s L_{smooth} + \gamma_r L_{regularization},$$

Testing

Given a input RGB image, the SMPL model (along with base depth) is computed by TrackNet and then depth offset map is estimated by ReconNet to get the final result.



Notes

- 24,000 frames with g.t. SMPL coefficients from their own dataset used for fine-tuning TrackNet
- 12,533 target-reference frames from their own generated dataset used for training ReconNet.
- 3,000 target-reference frames from Youtube used to finetune ReconNet. Fine-tuning improves both mid- and high- frequency shape details.
- Background for each complete video seq are taken from images in Places Dataset.
- $SSIM_{cs}$ is measured on the contrast domain and is more robust to misalignment (due to imperfect non-rigid motion estimation) and shading changes.
- Disadvantage of this method
    - “Complex clothing, faces and hairstyles are not estimated accurately mainly because they are difficult for photo-consistency based reconstruction.”

Takeaways

- This method is first step to full reconstruction
- Assumption: HMR need not work correctly for all frames
- Assumption: Base depth is accurately fitting the RGB input image
- Works only for tight clothing and exposed poses