Putting People in their Place: Monocular Regression of 3D People in Depth

Abstract

- Goal is to directly regress the pose and shape of all the people as well as **their relative depth.**
    - inferring the depth of a person is ambiguous without knowing their height.
- Propose new method, called BEV, **adds an Bird's-Eye-View representation to explicitly reason about depth**.
    - previous works estimated multiple people **by reasoning in the image plane.**
    - BEV reasons simultaneously about body centers in the image and in depth.
    - BEV is a single-shot method that is end-to-end differentiable.
        - It is follow-up study of ROMP (ICCV21)
- Exploit a 3D body model space that lets **BEV infers shapes from infants to adults.**
    - height varies with age
- **Create "Relative Human (RH)" dataset that includes age labels and relative depth relationships** between the people in the images.
- BEV outperforms existing methods on depth reasoning, child shape estimation and robustness to occlusion.

code: https://github.com/Arthur151/ROMP

dataset: https://github.com/Arthur151/Relative_Human

Overview

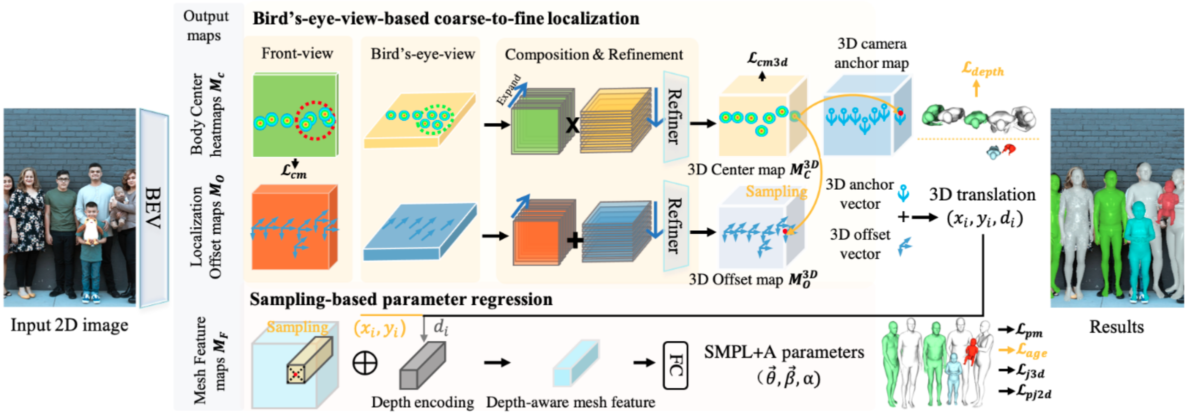

Figure 2. Overview. Given an RGB image, BEV first estimates the 3D translation of all people in the scene via compositing the front-view and the bird's-eye-view predictions. Then guided by the 3D translation, we sample the mesh feature of each person to regress their age-aware SMPL+A parameters. See Sec. 3.1 for details.