Novel-View Synthesis of Human Tourist Photos

Paper: https://ieeexplore.ieee.org/document/9706936

Authors: UoB PhD students & professors including Jon Freer & Hyung Jin Chang

▼ **Terminology:**

Image Registration: align 2 or more (part) images to the same scene ready for 3D reconstruction

Intrinsic camera properties: regardless of image, ie. blur & focal length of camera

Extrinsic camera properties: changes image-image, ie. camera position and angle

1. Objective

Given a known scene (generated from online images using computer imaging software COLMAP) and a tourist photo from the scene, let's reconstruct the scene with the tourist in novel positions and angles (viewpoints) whilst maintaining correct lighting & scaling conditions.

---

2. Approach

Stage 1: Human & scene digitisation (3D reconstruction)

In this stage:

- Reconstruct 3D mesh of the person and estimate their relative position to camera so they can be placed back into the scene later with (correct orientation and scales
- Reconstruct the 3D scene with a point cloud

1. Human detection (bounding box)

Mask RCNN locates all humans in input image (object detection with bounding boxes)

- segmentation from mask R-CNN was over-smoothed → poorly shaped human meshes

2. Human segmentation

- pre-trained DeepLab network
  - crisp segmentation without noise or over smoothing and better than using either network independently

3. Human mesh creation (’reconstruction’) to get human geometry

- pre-trained PIFu network

4. Scene point cloud creation

- COLMAP imaging software

5. Human depth estimation

- pre-trained MonoDepth2 network
  - depth estimations weren’t linear so needed to be aligned with point cloud coordinate system. Used a custom network to learn the scaling by comparing the depth of buildings estimated by MonoDepth2 to a ‘true depth’ generated by COLMAP computer imaging software

6. Image registration - align the 2 parts of the image and get camera properties ready for 3d scene reconstruction

- COLMAP software (again) which estimates camera intrinsic & extrinsic parameters
- Using the camera parameters, estimated depths & human positioning, we place the human mesh in the correct spot on the 3D reconstruction (correct scaling, position & orientation)

    - combines the 2 images into a scene by using estimated intrinsic and extrinsic camera properties, predicted depth & human position (pixel offset) with the human appropriately positioned and scaled in the 3D scene coordinate space

---

Stage 2: traditional renderer

In this stage we create a deep buffer (3d reconstruction database) using OpenGL (computer software) containing:

- pixel albedo
    - computer graphics approach to defining colour
    - ‘the colour (RGB) under 0 reflectivity/ in neutral lighting devoid of any lighting or shading affects’
- depth map
- binary segmentation map (allows neural renderer later to distinguish between pixels)
    - which pixels are human or non human

1. Colour in the mesh (geometry) of the human
    - also apply preprocessing in this step: normalising pixels & applying directional spherical harmonic lighting
2. Use custom lighting estimation network to obtain lighting coefficients for the human (where are reflections etc)
3. Use the previous scene (no human) reconstruction to compute a depth map and change the scene lighting based on depth
4. Combine scene (from point cloud) with human (from coloured mesh) to produce the deep buffer

---

Stage 3: neural renderer

Projects the deep buffer to 2D images

- Based of neural renderings in the wild model

Stage 4: relighting human

- Built dataset of 200 images of segmented humans with known lighting conditions
  - Augmented by relighting images using ‘learned lightprobes’ network which accurately changes lighting conditions
- Developed network based on ‘learned lightprobes’ to take in an image of a human and output coefficients representing the lighting
  - Use these lighting coefficients to relight from different angle when the human has been moved