Reading 2: Kinect Fusion and NeRF

Christopher Klammer

March 29, 2023

1 Kinect Fusion

1.1 Method

- The overall system comprises of surface measurement to produce vertex and surface normals to get a measurement structure, an update to the global scene using TSDF, surface prediction by using the global and local structures, and sensor pose estimation to find the alignment between the sensor measurement and the predicted surface.
- The surface measurement consists of using the camera perspective projection equation to map to real world points then applying a bilateral filter to apply smoothing where then each point is mapped to a 3D vertex, adjacent vertices are crossed in order to create a surface
- The paper suggests using a multi-scale representation by block averaging the finer representation then sub-sampling so long as the depth is not on outlier (spurious measurement that wouldn't contribute anyway)
- Mapping here using TSDF where positive distances are visible and negative distances are behind the surface, visible points points too far are truncated while non visible points are not measured, otherwise we just have a distance to the nearest surface point and will be averaged over time via a weight and both F_k and W_k are recurisvely filtered. The weight will be positive if it is seen before.
- Ray casting using a marching strategy for each depth pixel to predict the surface, they make a comment that if a point is lose to the surface, the gradient of the TSDF at the pixel can give the normal, interpolation is used to approximate the correct time of the zero crossing to be as accurate as possible.
- Sensor pose estimation uses a fast projective data association method and the point plane error metric to minimize the energy function between the normals of the surface and the predicted normals from the correspondence with the pose estimate where ICP is used for optimization.

2 NeRF

- NeRF represents continuous scenes with only basic MLP networks with a 5D input of location (x,y,z) and viewing direction (θ, ψ) with output (r, g, b) and volume density σ
- The weights are optimized to predict the volume density σ as a function of the location with 8 FC layers and outputs σ and a 256-D feature vector, then, the feature vector is concatenated with the viewing direction and then passed to one more FC layer to output the color c
- The continuous integral $C(r)$ is the expect color for the camera ray and must be estimated using quadrature and sampled by partitioning the near and far bounds of the integral by N evenly spaced bins which acts as a sufficient approximation and results in the MLP being evaluated at continuous positions
- Positional encoding of input coordinates helps to represent high frequency functions similar to transformers using a sinusoid for each component in x and d to create a higher dimensional representation to approximate the high frequency function
- Hierarchical sampling allows for efficient sampling by optimizing one coarse and one fine network, the coarse network uses the stratified sampling which is then passed into the fine sampling to take the best regions and boost importance to these areas
- L2 loss is used for the coarse and fine color prediction for a given ray r and 4096 rays, 64 coarse coordinates, and 128 fine coordinates with an ADAM optimizer