NeMo: Neural Mesh Models of Contrastive Features for Robust 3D Pose Estimation

Anonymous authors
Paper under double-blind review

Abstract

3D pose estimation is a challenging but important task in computer vision. In this work, we show that standard deep learning approaches to 3D pose estimation are not robust when objects are partially occluded or viewed from a previously unseen pose. Inspired by the robustness of generative vision models to partial occlusion, we propose to integrate deep neural networks with 3D generative representations of objects into a unified neural architecture that we term NeMo. In particular, NeMo learns a generative model of neural feature activations at each vertex on a dense 3D mesh. Using differentiable rendering we estimate the 3D object pose by minimizing the reconstruction error between NeMo and the feature representation of the target image. To avoid local optima in the reconstruction loss, we train the feature extractor to maximize the distance between the individual feature representations on the mesh using contrastive learning. Our extensive experiments on PASCAL3D+, occluded-PASCAL3D+ and ObjectNet3D show that NeMo is much more robust to partial occlusion and unseen pose compared to standard deep networks, while retaining competitive performance on regular data. Interestingly, our experiments also show that NeMo performs reasonably well even when the mesh representation only crudely approximates the true object geometry with a cuboid, hence revealing that the detailed 3D geometry is not needed for accurate 3D pose estimation. The code is publicly available at `https://github.com/Angtian/NeMo`.

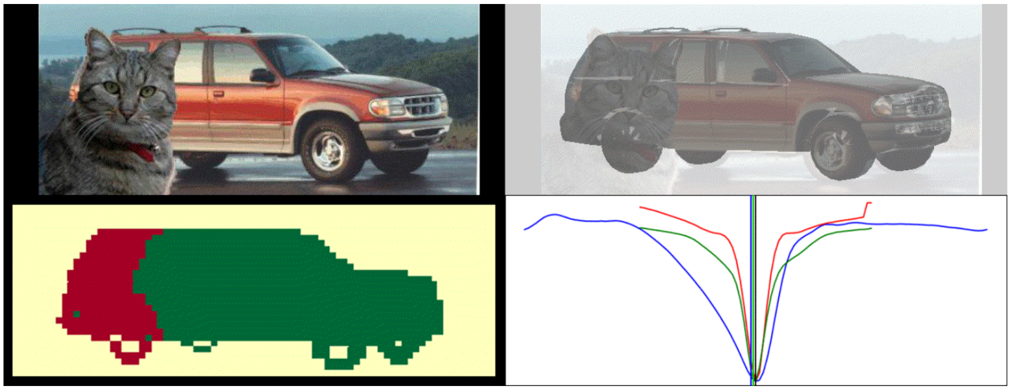

Figure 1: Qualitative results of NeMo on PASCAL3D+. Top-left: the input image; Top-right: A mesh superimposed on the input image in the predicted 3D pose. Bottom-left: The occluder localization result, where yellow is background, green is the non-occluded area of the object and red is the occluded area as predicted by NeMo. Bottom-right: The loss landscape for each individual camera parameter respectively. The colored vertical lines demonstrate the final prediction and the ground-truth parameter is at center of x-axis.