Abbreviated Writup
Evolving Dense Descriptors Towards Fine Manipulation

Bahaa Aldeeb - Fall 2021 - Directed Study

Overview:

Recent work [Florence & Manuelli] demonstrated the benefit of using Dense Descriptors (DenseNet) towards intra-category manipulation but an inspection of those descriptors show they might not be fit for more fine manipulation and pose estimation. That much was evident by the lack of details shown in the descriptors of the original work (Figure 1 - right). To remedy that issue this work improved the quality of descriptors by updating the backbone. Using deeplab's atrous convolution backbones yielded much finer results (Figure 1 left). While using DenseNet's training losses, we observed through quantitative experiments that the embedding still did not have sufficient information for pose estimation. In those experiments, we utilized the weighted-procrustes method to derive relative object poses and compared them to the true relative poses. That much can be also seen by observing that the features at the right side of the mugs' handles were not equivariant to the mug's change of pose. To try and remedy that issue we investigated the use of a new loss informed by object pose. Loosely inspired by NOCS that loss associates intra-category object surfaces by scaling and centering said objects and using cosine similarity to build correspondences between their point clouds. Figure 2 below depicts the generated associations. The results, though visually more consistent with pose, yielded a collapse of part of the descriptor space and consequently, no pose estimation benefit. This work was aimed towards generalizable affordance detection, it was benched with the end of the Fall term.

Selecting the Backbone

The original work used a backbone that downsized the image then deconvolved (upsized) which placed a burden on the network to learn how to re-generate details. The objective of preserving detail has been the focus of work on segmentation. To relieve networks from having to regenerate detail, some networks (UNet, Pyramidal) used residual connections between different granularity levels of a convolution pyramid [Lin et al.]. Residual connections, while helpful, generated results similar to those of the original work. Atrous methods [Liang-Chieh Chen et al.] relied on varying the shape of the convolutional filter while preserving the original size of the image through the pipeline. In doing so, Atrous methods produced very crisp detail while maintaining the ability to learn distributions rather than only masks. The figure below shows results from both networks. Both networks are implemented in pytorch which facilitated their integration into our work.

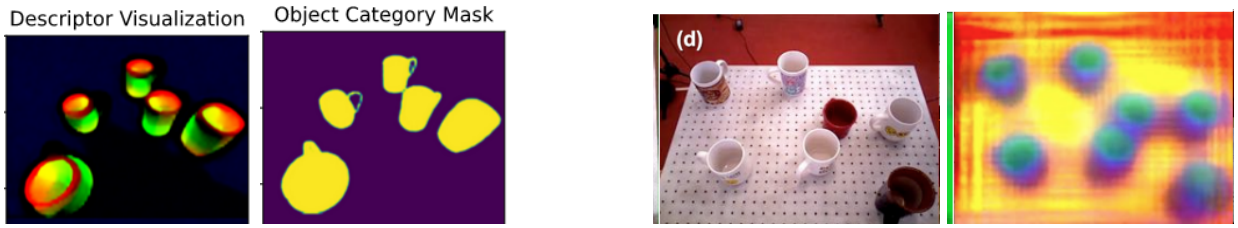


Figure 1: To the right is a sample descriptor from the original work [Florence & Manuelli]. To the left are a descriptor and mask produced by the network we developed.

Introducing Intra-Category Loss

We noticed from our visualizations that the distribution shown over an area of a mug does not consistently vary (A mug handle sometimes has different descriptors based on its position in the image). We care for them to vary consistently within one object category so that we could use them to generalize our labeled data. For the sake of experimentation, we introduced a new loss to explicitly encourage distributions to vary as we desire.

The data used to run experiments consists of mugs and their original pose. The original pose of a mug is designed to align handles and thus respects the intra-category similarities between objects. Inspired by work on pose and size estimation [Wang et al.], we projected each detected object into 3D, scaled them into a unit cube, and positioned them in their original pose. The center of an object becomes the origin which allows us to associate pixels of objects with different shapes using cosine similarity as a measure. The images below show associated pixels in two very differently shaped mugs.

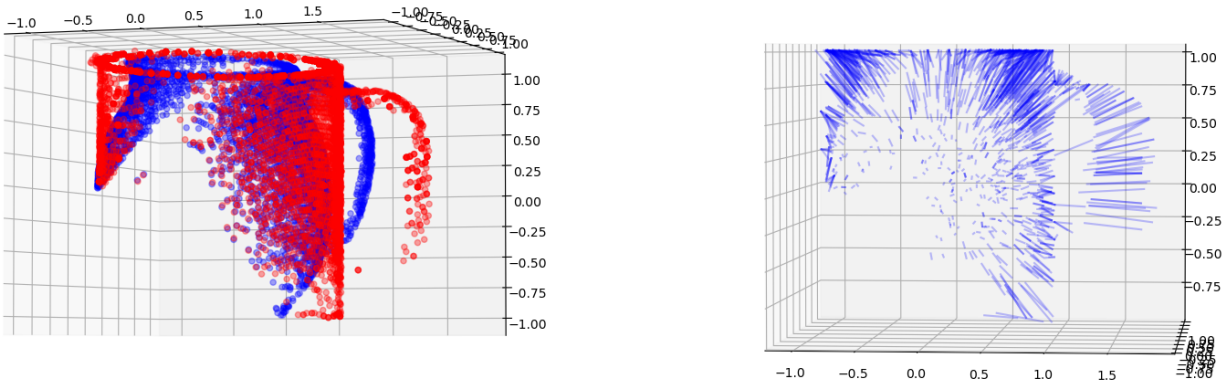
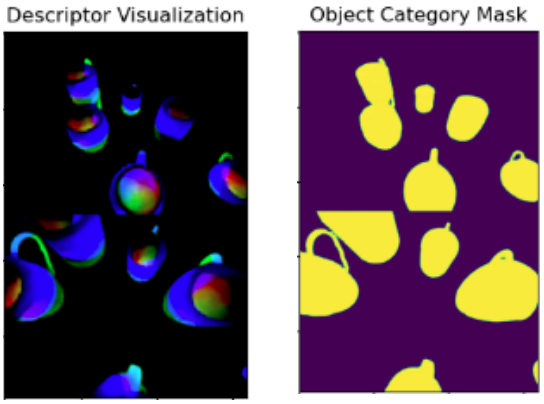


Figure 2: To the left is the projection of the RGB-D image into 3D. To the left are vectors showing associated pixels using the cosine similarity of each point relative to the center of the objects.

Deriving those associations allows us to encourage pixels on different instances to have similar features to look similar. The vision is to possibly have all handles similarly distributed in descriptor space.



The result of using such loss encouraged the intended outcome but only at the cost of collapsing part of our distribution as evident from the images to the right. One can see that almost all handles are similarly colored but there is no longer a sufficient variation in descriptor space to allow us to easily derive pose information.