CSCI 5980-DeepRob

Group 4

One-Shot Imitation Learning Reimplementation of DOME paper

Tzu-Hsien Lee, lee04484@umn.edu
Fidan Mahmudova, mahmu059@umn.edu
Rammesh Adhav Saravanan, sarav060@umn.edu

Introduction:

This poster introduces a one-shot imitation learning approach designed to enable a robot to replicate a demonstrated task based on a single image-based example. Building upon the DOME [1] framework, our method allows the robot to first determine the object-relative pose of its end-effector from the given demonstration image, and then reproduce the demonstrated joint angles to complete the task.

Input-output:

During inference, the system takes in two images of the same size (64×64×3): a live view showing what the robot currently sees, and a goal (bottleneck) image representing the desired final state. Using these two inputs, the model produces five values that describe how to move and rotate the robot's end-effector— small adjustments along the x, y, and z directions, plus two for rotation along z axis.

Framework details:

The overall training framework comprises two main components as illustrated in Network Architecture section: a segmentation network (Stage I) followed by a learned visual servoing network (Stage II). The segmentation network adopts a U-Net architecture, applying Tiling and FiLM layers at the bottleneck image to enhance feature extraction. The learned visual servoing network then uses Siamese CNNs, operating on these segmented images to output positional and rotational adjustments for the robot's end-effector.

Dataset information:

A total of 120 trajectory episodes were recorded using a Kinova arm in the ROS Gazebo simulation. The dataset includes 1,300 pairs of live and bottleneck images, along with corresponding fine adjustment values required for the arm to move from the bottleneck position to the next live image position. The split for training and testing was 95% and 5% , accordingly.

Evaluation: What did you experiment on?

We conducted experiments comparing two segmentation models trained to identify objects in a scene: a base model trained only on images containing a mug, and a one-shot model trained using just a single example trajectory that includes cans alongside the mug. We evaluated their performance using Intersection over Union (IoU) to see how well each model could segment objects (Figure 1).

Another evaluation metrics was the measurement of success of translation and rotation actions respectively taken by the models mentioned before. (Figure 2).

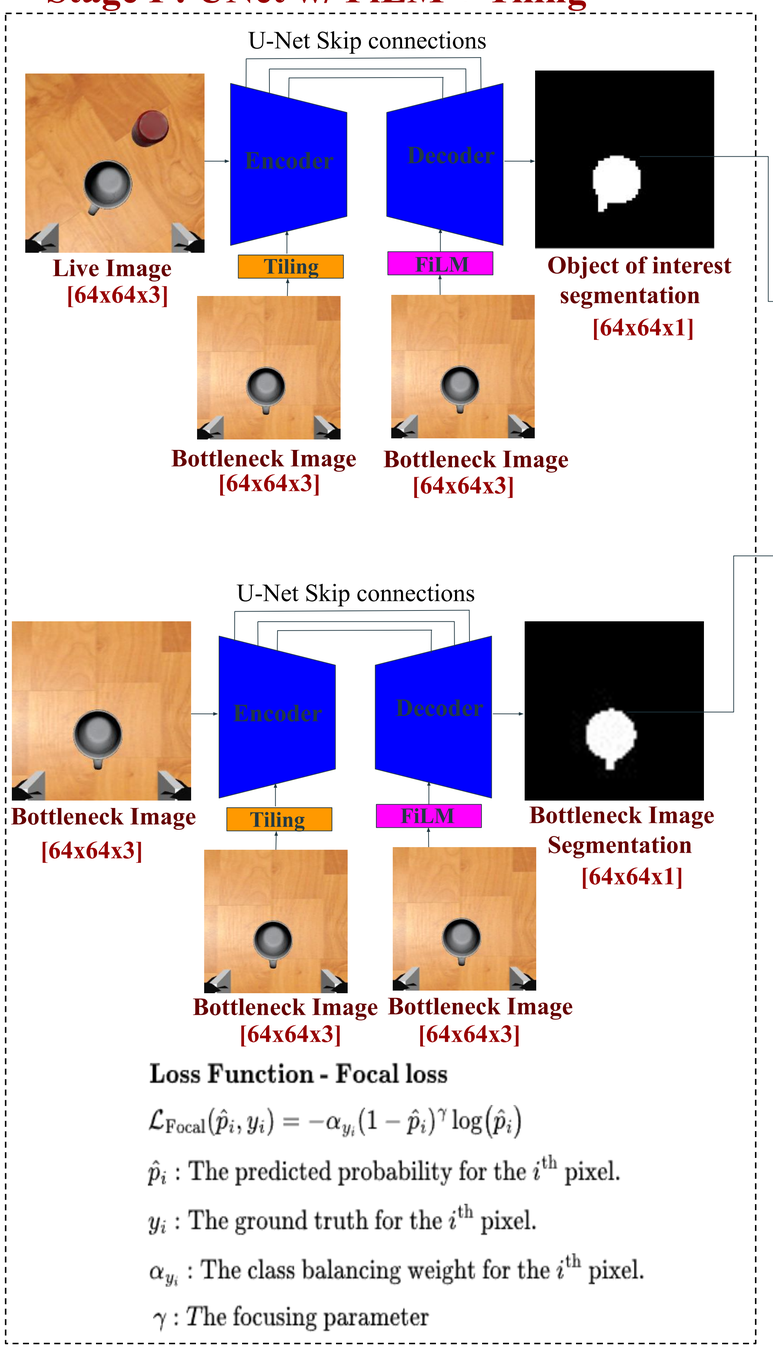
References:

[1] E. Valassakis, G. Papagiannis, N. Di Palo, and E. Johns, "Demonstrate Once, Imitate Immediately (DOME)," *Proc. IEEE/RSJ IROS*, 2022.

"One-Shot Imitation Learning Using Visual Demonstrations."

1. Network Design

Stage I : UNet w/ FiLM + Tiling



Loss Function - Focal loss

$$\mathcal{L}_{\text{Focal}}(\hat{p}_i, y_i) = -\alpha_{y_i}(1-\hat{p}_i)^{\gamma}\log(\hat{p}_i)$$

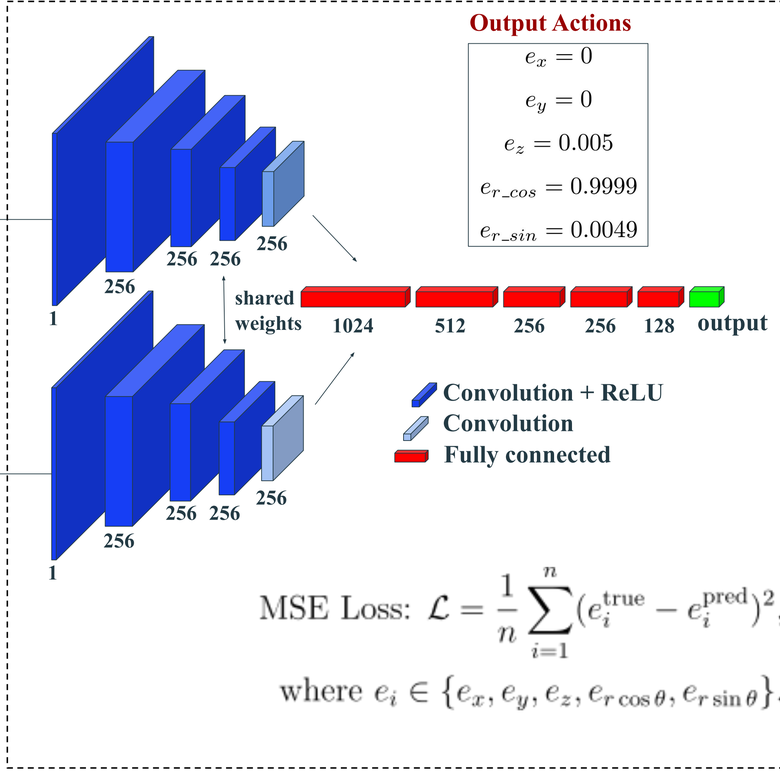
$\hat{p}_i$: The predicted probability for the i^{th} pixel.

y_i : The ground truth for the i^{th} pixel.

α_{y_i} : The class balancing weight for the i^{th} pixel.

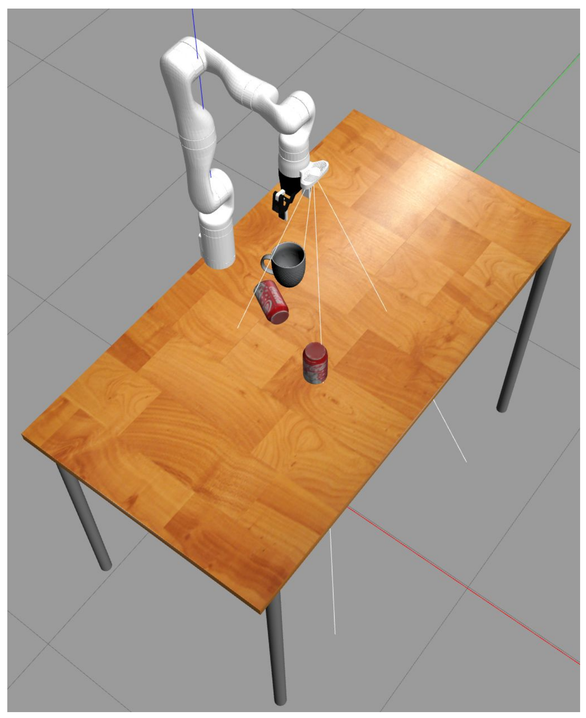
γ : The focusing parameter

Stage II : Siamese Network



$$\text{MSE Loss: } \mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(e_i^{\text{true}} - e_i^{\text{pred}})^2,$$

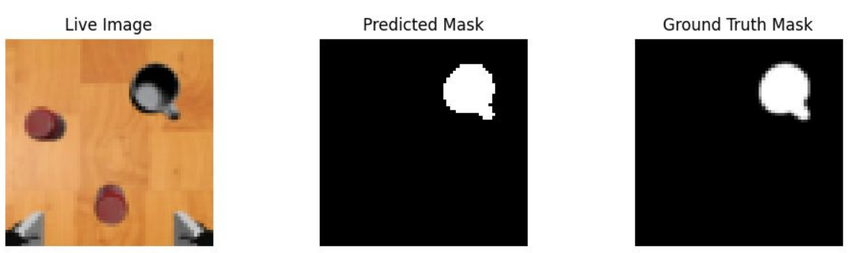
$$\text{where } e_i \in \{e_x, e_y, e_z, e_{r\cos\theta}, e_{r\sin\theta}\}.$$



2. Evaluation Results

Base Model

One-Shot Model



	Only Mug	*Mug With Cans*
Base Model	0.919	0.756
One-Shot Model	0.916	0.903

Figure 1. Modified Unet architecture IOU evaluation

		Only mug	*Mug with cans*
Base Model	***Translation Success Rate***	80 %	70 %
	Rotation Success Rate	20 %	10 %
One-Shot Model	***Translation Success Rate***	80 %	60 %
	Rotation Success Rate	20 %	50 %

Figure 2. Complete Model Success Evaluation

computer science & engineering

MINNESOTA ROBOTICS INSTITUTE
UNIVERSITY OF MINNESOTA