COMS 4733 **Ruogu Li** **rl3318**

Homework 4

Notes you want the TAs to consider when grading.

Problem 2

(a).

Why do we use get gaussian scoremap to generate the affordance target, instead of a one-hot pixel image:

The reason for using gaussian scoremap instead of one-hot pixel map is that the latter can only represent a discrete grasping point, typically with a pixel value of 1 at the grasping point and 0 elsewhere, which is sensitive to position estimation errors and may lead to overfitting of the model. In contrast, gaussian scoremap is smoother, can tolerate a certain range of grasping point deviation, and better represents possible grasping areas. Additionally, gaussian scoremap can better calculate the loss function and provide more gradient information, which is beneficial for model training and optimization.

(b).

The self.aug_pipeline in the AugumentedDataset class applies two random geometric transformations to the input RGB images:

1. Translation: The image is shifted along the x and y axes within a range of -20% to 20%.
2. Rotation: The image is rotated within a range of -11.25 to 11.25 degrees. (angle_delta equals 180/8 = 22.5 degrees)

These transformations are applied with a 70% probability to increase the dataset's diversity and improve the model's generalization ability. The pipeline also updates the keypoints (left and right finger positions) to keep them consistent with the augmented image.

(d).

Train Loss: 0.0013
Test Loss: 0.0011

Fig 1: training_vis

(f).

Success Rate: 0.800000011920929

Link: https://drive.google.com/file/d/1EhgZekhTooANFht_67TMC8X1N-tZlhwy/view?usp=share_link

Fig:

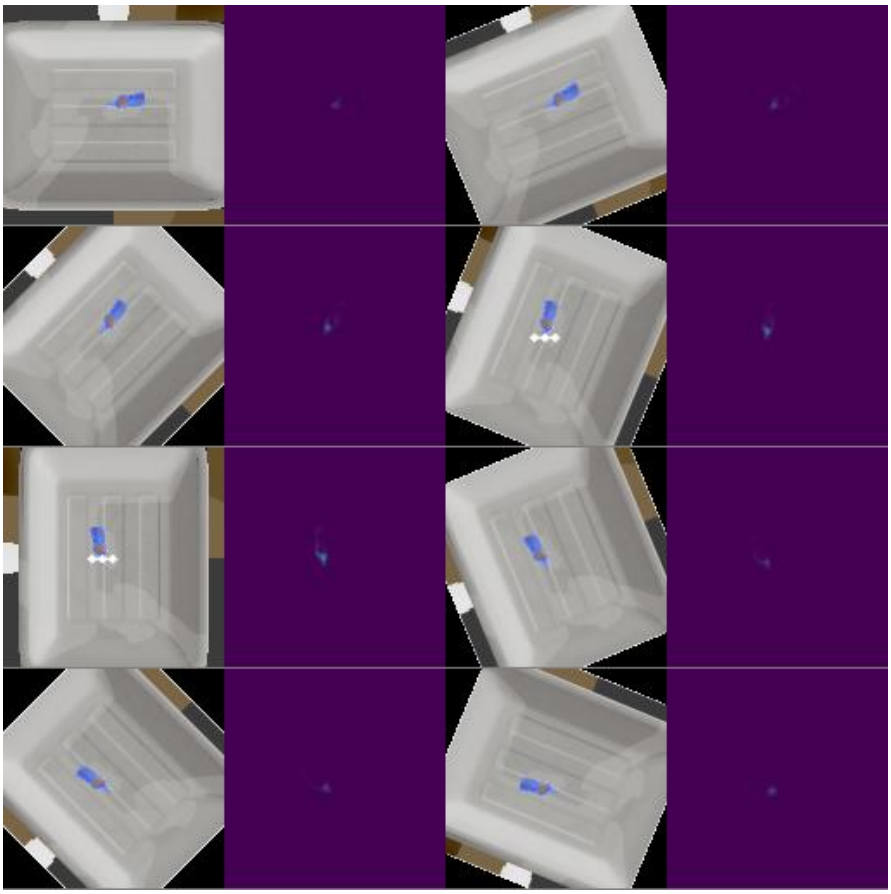

Fig 2: YcbMustardBottle_1

(g).

Success Rate: 0.6899999976158142

URL: https://drive.google.com/file/d/1ZPj1gR2xyQ4rQ4l8H6A2-5PaauiXajLH/view?usp=share_link

Fig:

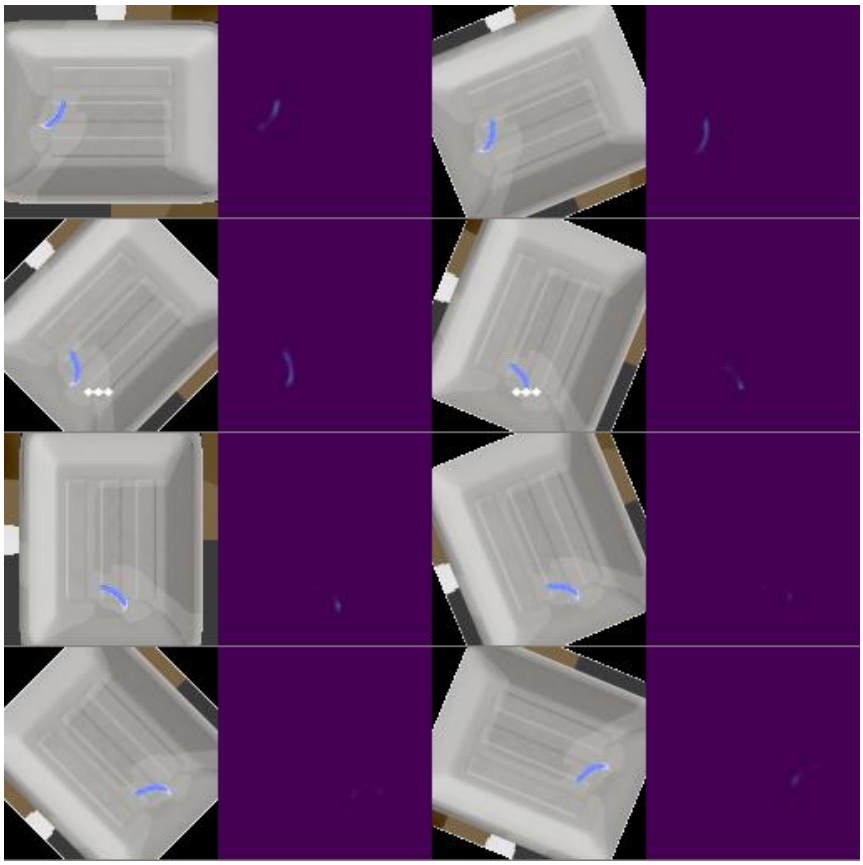

Fig 3: YcbBanana_8

Describe:

The image might be different from the visualizations in the previous part because the model's performance on novel objects may vary from its performance on the training set. When the model encounters new objects in the testing set, it might be less accurate in predicting grasp points and angles, leading to differences in the visualizations compared to the results on the training set. However, similarities may come from the model's ability to learn general grasp affordances that can still apply to the novel objects in the testing set. This means that, in some cases, the model may still be able to generate visualizations similar to those in the training set because it can capture fundamental grasping patterns applicable to various objects.

(h).

Objects Left: 2

Seed 2

URL: https://drive.google.com/file/d/1f4GxpZBauvaMUphkarWSNs8pXHAaAC5U/view?usp=share_link

(i).

1. Grasp representation: The Gaussian scoremap offers a continuous and informative representation for grasp affordances, unlike one-hot pixel images.

This flexibility enables the model to acquire more general grasp affordances, applicable to various objects.

2. Data augmentation: Techniques such as rotation and translation effectively expand the training dataset by creating new image variations. This increased sample diversity enhances the model's robustness to different object poses and viewpoints, resulting in improved performance on unseen objects.

3. Transfer learning: Utilizing a pre-trained backbone network allows the model to benefit from knowledge gained on a large dataset. This equips the model with valuable object recognition and scene understanding features, facilitating better generalization to unfamiliar objects.

4. End-to-end training: The model focuses on directly optimizing the grasp prediction task, enabling parameter fine-tuning and improved performance on both familiar and unfamiliar objects.

Problem 3

(a).

Success rate: 0.800000011920929

URL: https://drive.google.com/file/d/19PT7oXBY94F_LdyxL8B6NQVsO-eeRc5l/view?usp=share_link

(b).

Success rate: 0.8500000238418579

URL: https://drive.google.com/file/d/1xFYDe78J2yRPa8ScZ_WQPzO-AYXsNmX-/view?usp=share_link

(c).

Objects Left:2

Seed 2

URL: https://drive.google.com/file/d/1FPrHcsXpgSS_56GBveTXybCj8WZsZBiw/view?usp=share_link

(d).

In Problem 3, the performance difference during each evaluation run compared to Problem 2 comes from the additional code which enables the model to suppress past actions and select the next-best action. This is achieved by maintaining a list of past actions and applying a Gaussian suppression map to the predicted output, which prevents the model from selecting the same failed action again.

In Problem 2, the model only attempts to grasp an object once without considering the past failed actions. Therefore, it may not perform well on challenging objects or situations. On the other hand, in Problem 3, the model has the ability to learn from its past failed actions by suppressing them and selecting a new, potentially better action. This iterative process of attempting to grasp an object multiple times with different actions can lead to improved performance in grasping both seen and unseen objects.

Problem 4

(a).

Train Loss:0.0125

Test Loss:0.0324

Fig:

Fig 4: training_vis

(b).

Success Rate: 0.13333334028720856

URL: https://drive.google.com/file/d/1ouCilF6rEWOg3RvIieAsSd6npXekZFMs/view?usp=share_link

Fig:

Fig 5: YcbMustardBottle_2

(c).

Objects Left:12

Seed 2

URL: https://drive.google.com/file/d/1vIuJ4m73lDABbxFQSu5q-AECa3QwieNH/view?usp=share_link

Explain:

1. Structural differences: Action regression directly predicts specific actions (e.g., grasping position and angle), while Visual Affordance models object surfaces, offering richer context and adaptability to shape and posture changes.
2. Objectives: Visual Affordance learns to predict score maps for actionable areas, capturing surface features more effectively than Action regression's direct action prediction, resulting in better real-world performance.
3. Generalization: Visual Affordance's surface feature modeling leads to stronger generalization, whereas Action regression may suffer from insufficient training data or overfitting, causing weaker generalization in practice.
4. Robustness: Visual Affordance better handles visual condition changes, such as

lighting and occlusions. In contrast, Action regression may underperform due to these factors, affecting prediction performance.

In summary, Visual Affordance outperforms Action regression in structural differences, objectives, generalization, and robustness, resulting in superior practical performance.