10/25

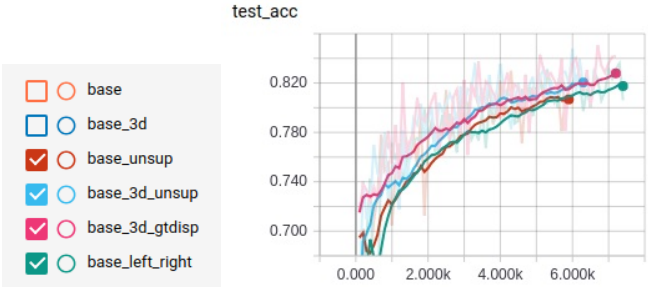
Project: Do unsupervised low-level depth and optical flow help supervised high-level visual understanding?
- **Context**: Try to understand how potentially unsupervised learned depth and optical flow can help object recognition and detection. Either Use ground truth depth and flow or use stereo video pairs (4 frames as input) from Wang Yang's algorithm to produce the truly unsupervised signal. Then combine RGB, depth, and flow as input signals to check. Conduct experiments on MLT, VDrift, KITTI, and CityScape dataset.
- **Owner**: Yi Yang, Yang Wang, Liang Zhao
- **Progress**: 1. Submit 3D fine-grained paper to Arxiv. 2. Wrap up code

10/18

- **Owner**: Yi Yang, Yang Wang, Liang Zhao
- **Progress**: 1. Train unsupervised flow on CityScape dataset. 2. Use unsupervised learned optical flow as extra input, the performance does not improve yet.

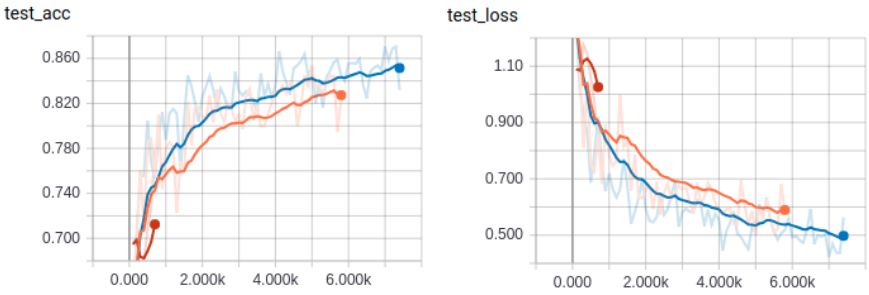
10/11

- **Progress**: Verify that unsupervised disparity indeed helps semantic image segmentation on CityScape dataset. What is more interesting is that, using both left right image as input does not help segmentation results.
- **Todo**: 1. Testing accuracy v.s. #train data. 2. Optical flow as extra input 3.



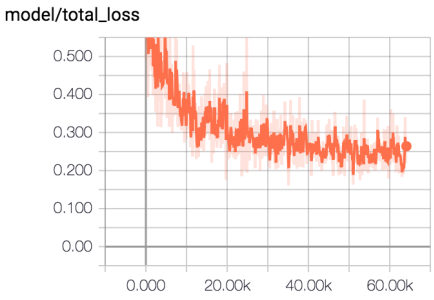
10/4

- **Progress**: Train semantic segmentation on CityScape dataset, using existing disparity helps.



9/27

- **Progress**: Fix bugs in unsupervised training data provider for Cityscape, now training stereo disparity estimation model, EPE looks reasonable now.



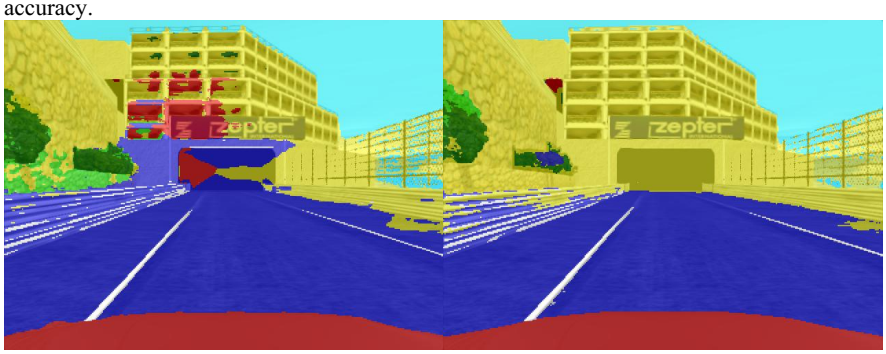
Project 2: Improved Annotation for 3D fine-grained pose dataset
- **Context**: Use image segmentation to refine the 3D pose annotation. Results look promising.
- **Owner**: Yi Yang, Feng Zhou
- **Progress**: Finish producing annotation for both StanfordCars and FGVC-Aircraft data. Working on submitting the camera-ready submission.

9/20

- **Progress**: Finish unsupervised training data provider for CityScape, writing testing data provider.
- **Progress**: Finish producing annotation for StanfordCars training data and FGVC-Aricraft training/testing data, now continuing on StanfordCars testing data.

9/13

- **Progress**: Train semantic segmentation on VDrift dataset, using RGB 89% testing accuracy, using RGBD 91% testing accuracy.



- **Progress**: Finish producing annotation for StanfordCars and FGVC-Aricraft training data, now continuing on testing data.

9/6

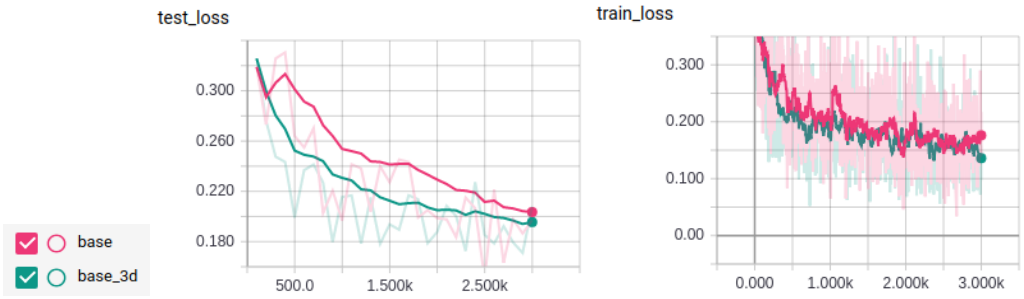
- **Progress**: Organize code for KITTI detection, able to run on AI cluster, download all required datasets to AI cluster.
- **Progress**: Continue working on the writing of 3D fine-grained object pose estimation dataset report.

8/30

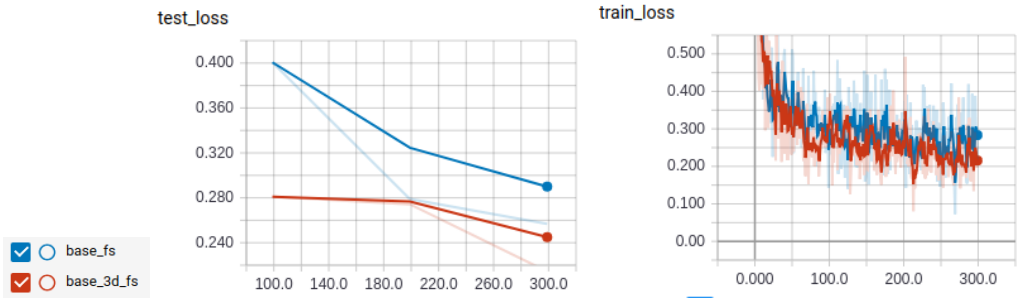
- **Progress**: Downloaded all CityScape video dataset, waiting for Wang yang's unsupervised code to be ready.
- **Progress**: Paper accepted at ECCV workshop. Continue working on the writing of 3D fine-grained object pose estimation for CVPR.

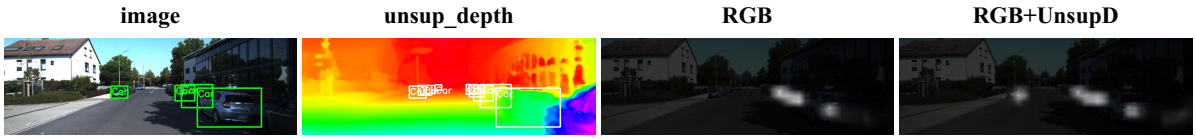
8/23

- **Progress**: Verified: adding unsupervised depth (from Wang Yang) is helpful in lowering detection loss.
- Compare two experimental settings:
    1. Kitti car detection. Use 5984 training images and 1497 testing images, train 3000 iterations, same VGG-type network structure. Adding unsupervised depth decreases the detection testing loss.



2. Kitti car few shot detection. Use 598 training images and 1497 testing images, train 300 iterations, same VGG-type network structure. Adding unsupervised depth even more significantly decreases the detection testing loss.



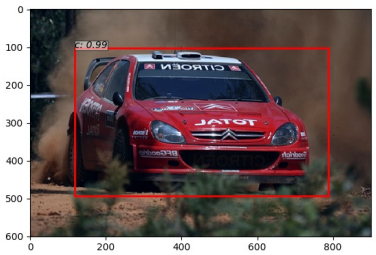


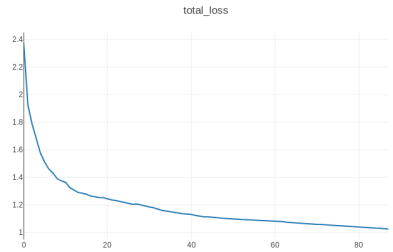
8/16

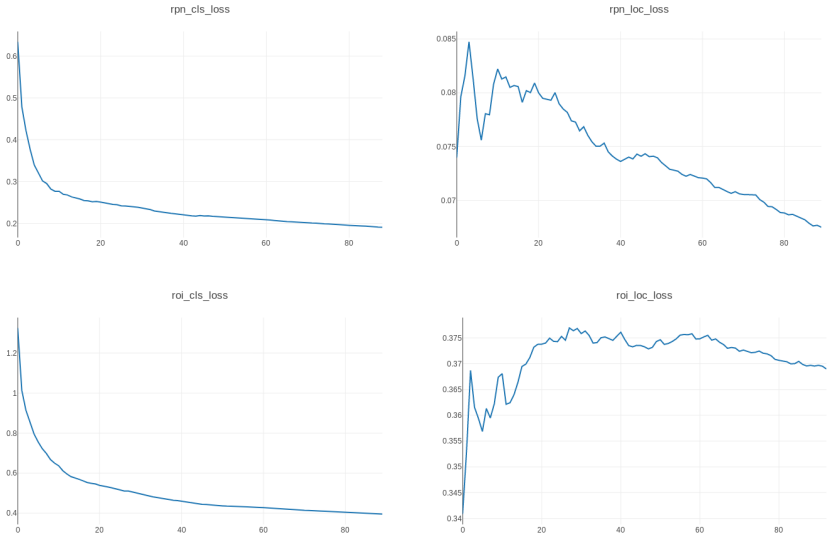
- **Progress**:
  - Manage to train a preliminary RGB based object detection on KITTI car detection.
  - Adding unsupervised depth (from Wang Yang) seems helpful in lowering detection loss but needs further verification.
  - So far both models in RGB and RGBD does not overfit on testing data.
- **Todo**
  - Will enlarge the model complexity to reach the extreme for each model.
  - Will make detection results more through so can be evaluated using mAP.

8/9

- After studying the dataset distribution on KITTI, decide to work with KITTI detection problem instead of VQA problem. The main reason is the class is dominated by cars which brings trouble for the model to learn a good generalization on locations to recognize other objects.
- Modify Faster RCNN code it make it runnable on PyTorch 0.4.0. Test the network training on PASCAL VOC 2007 dataset. The prediction looks reasonable and loss converges smoothly. At the stage of writing KITTI data provider for Faster RCNN code.

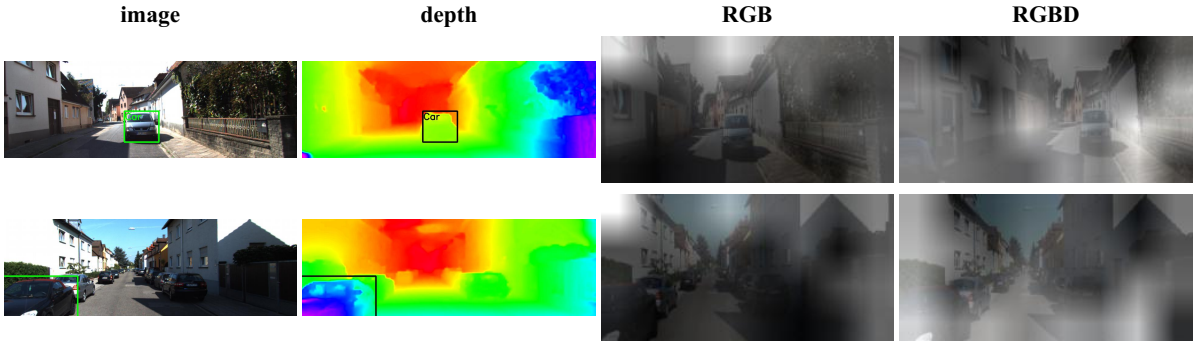






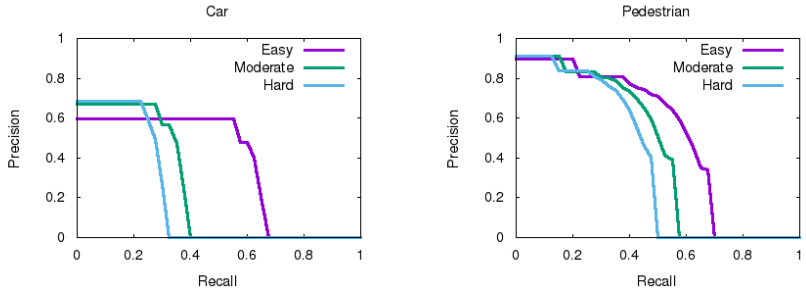
8/2

- Involve top-down location guidance on KITTI attention VQA. With top-down location guidance, RGB recognition accuracy improve from 81% to 86% on testing, RGBD recognition accuracy improve from 81% to 87% on testing. There is so far 1% improvemeent adding unsupervised depth.
- However, both attention maps look not meaniful, both models need more analysis.
- The attention map also suggest that the current experimental setting needs further change. This is because the car class is dominate (75%). Although one can improve to 81% or even 87% accuracy, we shall setup a new experiment on this dataset, such as adding background class.



7/26

- Setup RGB baseline for KITTI object detection. Use the Faster-RCNN model pretrained on Microsoft COCO to detect KITTI car and pedestrians. The model has not been finetuned yet. And there may also be bugs there.



- Setup baseline for KITTI attention VQA. Use the previous attention model on KITTI car and pedestrian VQA. Still under training. So far, using both RGB and RGBD gets 81% accuracy on testing. Adding unsupervisedly learned depth shows no improvement yet. I will need more time to analyze the results and get to more sophisticated model.

7/19

- Continue working on the writing of 3D fine-grained object pose estimation with Feng Zhou for another ECCV workshop. Use image segmentation to refine the initial pose annotation by our annotators. Results look promising.
- Left column is original pose annotation. Right column is refined pose annotation.

7/12

- Finish the writing of zero-shot transfer VQA paper writing with Yuanpeng and submit to ECCV workshop
- Work on the writing of 3D fine-grained object pose estimation with Feng Zhou for another ECCV workshop
- Work with Wang Yang on extracting optical flow, depth and moving object masks on the KITTI object detection dataset.

7/5

- Finish downloading the KITTI object detection dataset with stereo and 3 consecutive frames.
- Continue helping Yuanpeng on zero-shot transfer VQA paper writing.

6/28

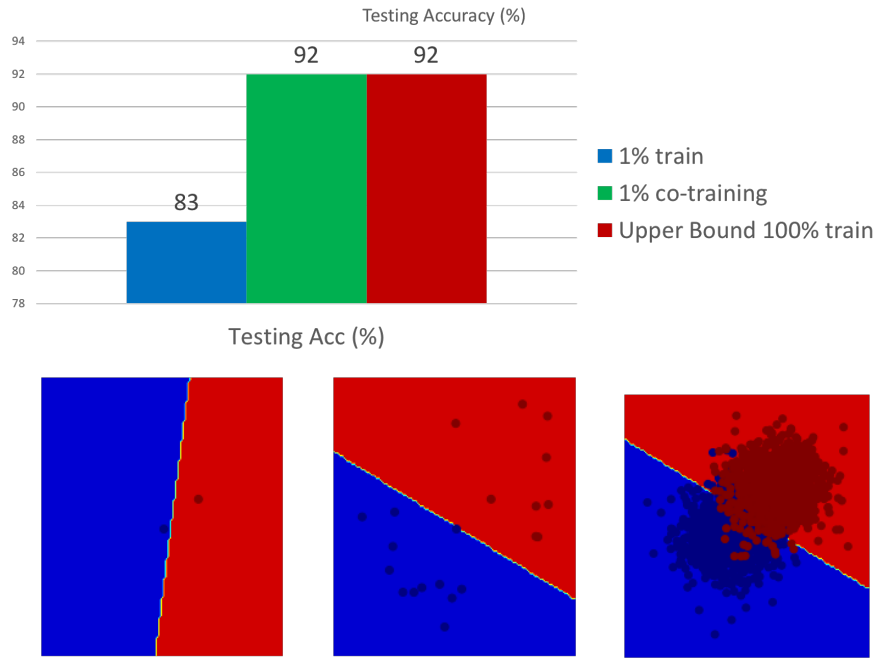
- Start working on KITTI object detection with depth and optical flow. So far, still preparing the data for the training and validation.
- Publish feedback neural networks on TPAMI and organize the code.
- Helping Yuanpeng on zero-shot transfer VQA paper writing.

6/21

- One week off

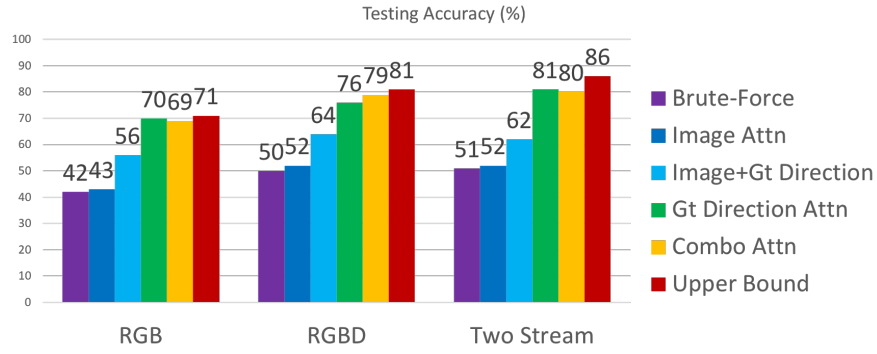
6/14

- Finish the co-training pipeline, test it on a simple 2-dimensional Gaussian Mixture data, find it indeed is helpful.
- More specifically, I generate 200 training data for 2 classes with each class 100 training. I only annotate 1 data for each class, so the supervision is very limited. The baseline (blue) is the testing accuracy using only 1 data per class to supervise the classifier. The upper bound (red) is the testing accuracy using all 200 data per class to supervise the classifier. The co-training (green) use 1 labeled data per class and 99 unlabeled data per class to train.



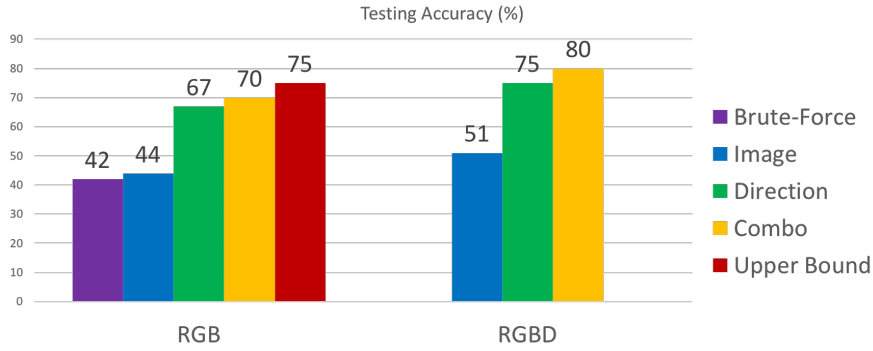
6/7

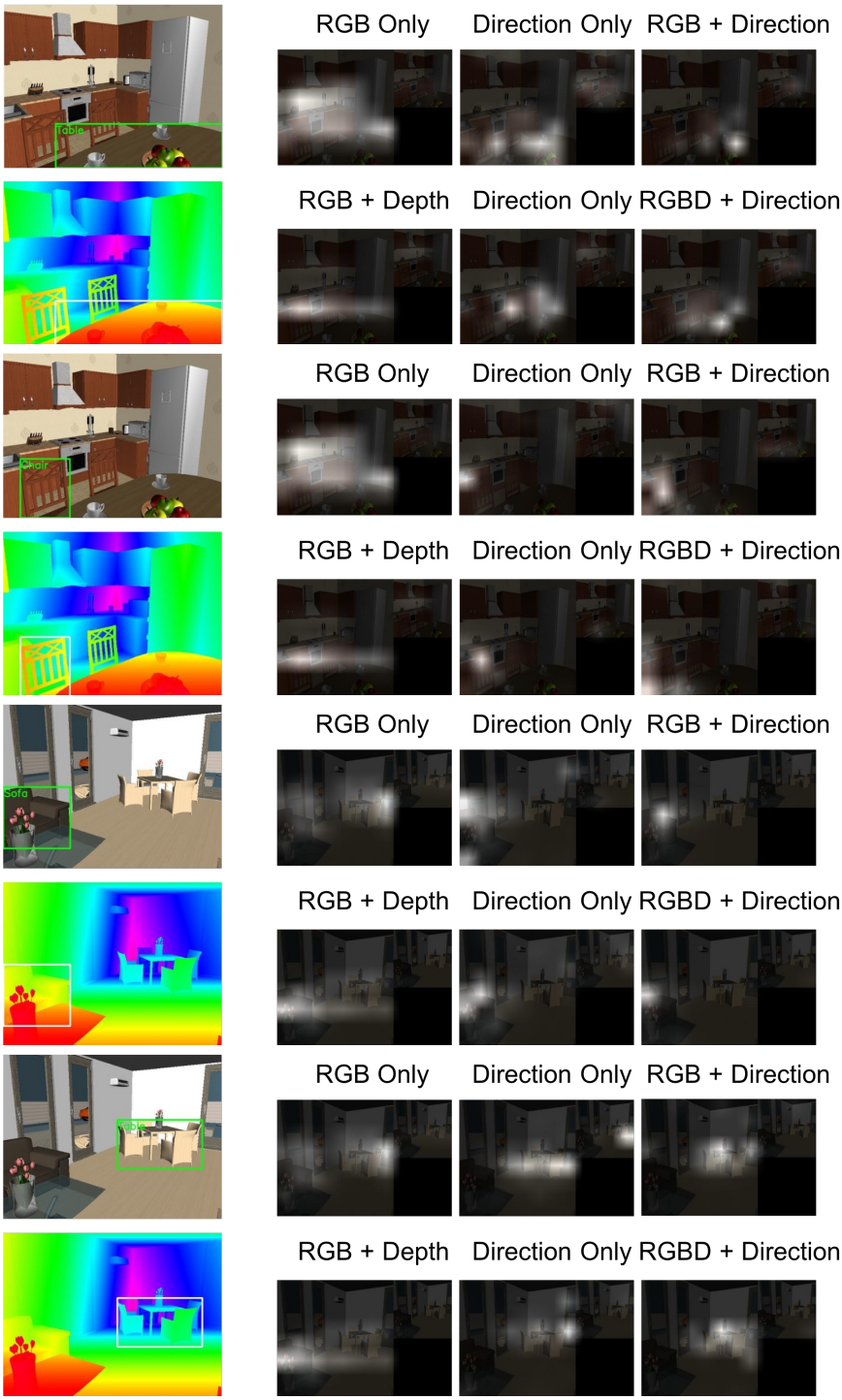
- Adding more baseline comparisons according to last week's discussion. Starting to study co-training for weakly supervised multi-class classification. In order to do co-training, the networks need to have two separate branches with each branch has the ability for classification. Hence I modify the network structure from previously using a CNN based on 4 channel input (RGBD) to two-stream CNN with one stream based on 3 channel RGB and another stream based on 1 channel depth.
- It seems the two-stream model works even better.

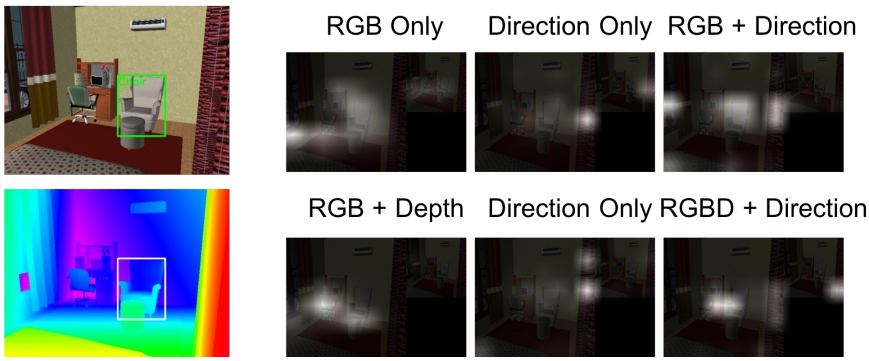


5/31

- Summarize the soft-attention model experiments on MLT dataset.
- Overall, Depth helps RGB for ~10% testing accuracy which is significant.
- Below is a visualization of quantitative results and qualitative results:


Testing Accuracy (%)
90
80
70
60
50
40
30
20
10
0
42
44
67
70
75
51
75
80
RGB
RGBD
Brute-Force
Image
Direction
Combo
Upper Bound

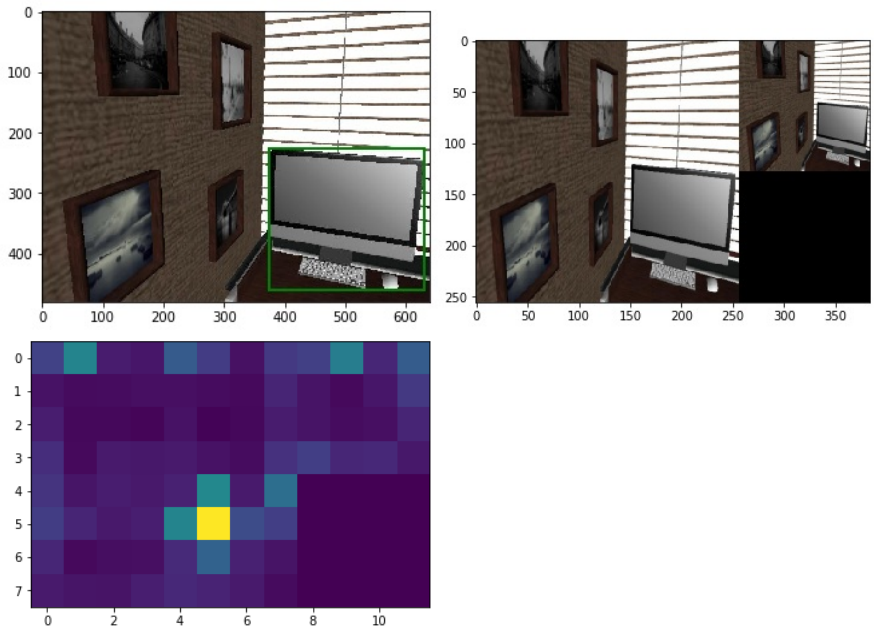

RGB Only
Direction Only
RGB + Direction
RGB + Depth
Direction Only
RGBD + Direction
RGB Only
Direction Only
RGB + Direction
RGB + Depth
Direction Only
RGBD + Direction
RGB Only
Direction Only
RGB + Direction
RGB + Depth
Direction Only
RGBD + Direction
RGB Only
Direction Only
RGB + Direction
RGB + Depth
Direction Only
RGBD + Direction



5/28

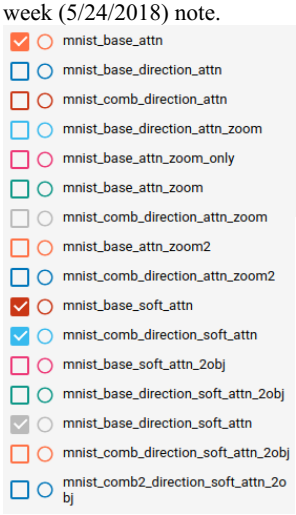
- Working on soft-attention model and switch back to MLT dataset because Mnist dataset is too simple.
- On MLT dataset, there are two main promising conclusions:
    1. Adding depth significantly helps image recognition. For example, the testing recognition accuracy increases from 42% to 48% by adding depth. And after adding top-down direction signal, the testing accuracy further increases from 48% to 72% which is about 30% total absolute improvement to the baseline RGB only.
    2. Adding top-down direction as keyword to obtain attention significantly helps image recognition. For example, the testing recognition accuracy increases from 48% to 61% by adding top-down direction on RGB image.
- In total, the depth and direction keyword can significantly improve recognition accuracy from 42% to 72%, increasing 30%, and the training actually has not converged yet. And all curves haven't converged yet.

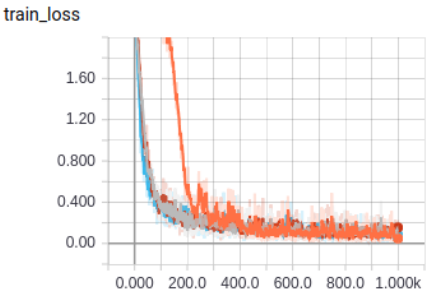
- An illustration of the soft attention model

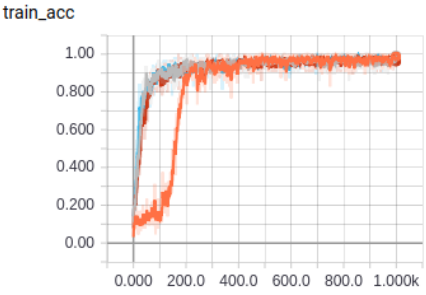


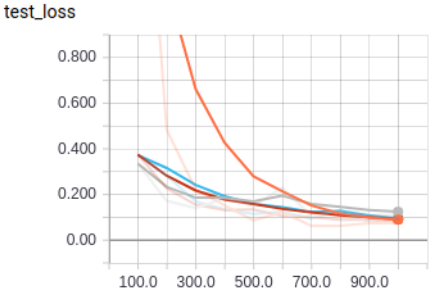
5/25

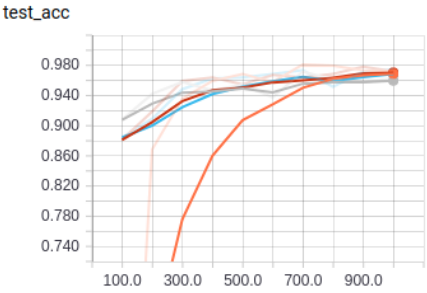
- Conduct comparison between hard attention model and soft attention model on Mnist dataset. Conclusion: using soft attention model performs much more smoother and faster convergence compared to previous hard attention (spatial transformer networks). The oracle performance of soft attention model may be slightly worse than the best hard attention model on recognition, however, in reality training converges much smoother.
- Here I show the training convergence using the soft attention model and a baseline hard attention model (orange) on Mnist dataset. One can see the three (red, cyan, gray) soft-attention curves all converge much faster than the (orange) hard-attention (spatial transformer networks) curve. For more hard attention model performances, please see the last week (5/24/2018) note.







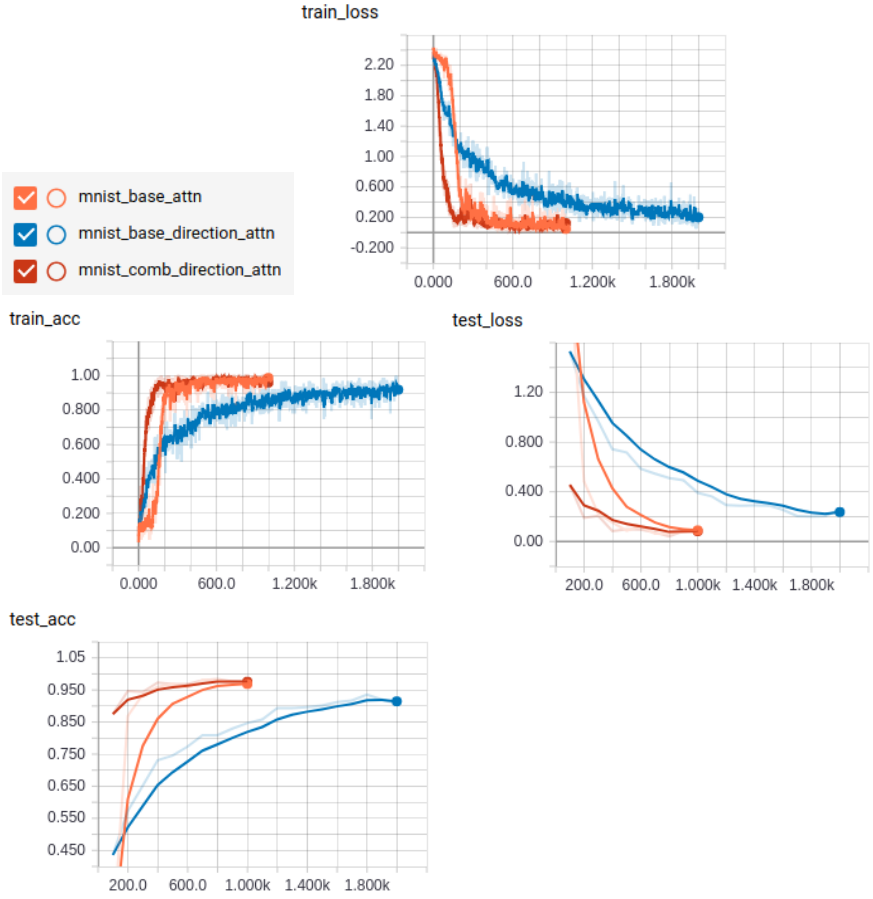




- On Mnist two object dataset, using the ground truth direction lead to the best convergence which now makes a lot of sense.

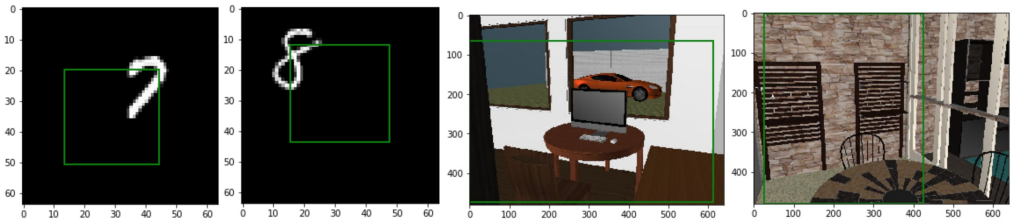
5/24

- The attention model (spatial transformer networks) can work now on Mnist dataset, when Mnist digits are randomly uniformly located in the image. Both image-based attention and word-based attention can provide reasonable attention. Jointly train them can achieve even better and faster convergence on learning classifiers and attention models.



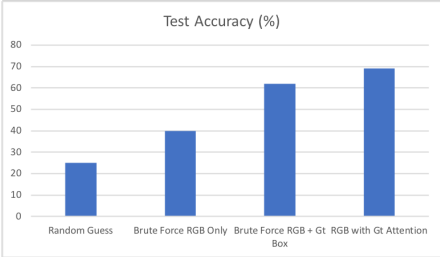
5/17

- The attention model (spatial transformer networks) does not work on MLT dataset so far.
- I find the model struggling in finding the correct object location, and is very sensitive to initialization.
- Debug attention networks on Mnist dataset-
  and find it even fails on simple Mnist data. Below is an illustration



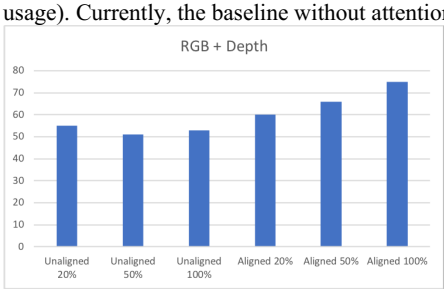
5/10

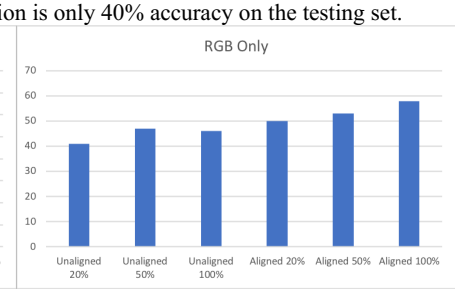
- Start to involve attention models to recognize objects. Previously we manually crop the image patch based on the bounding box. Now we use spatial transformer networks to crop the image based on the bounding box.
- To make training/testing more stable, now use 10742 training images and 4605 testing images, 6 classes on MLT dataset.
- Performance summarization: Random guess is 25%, brute-forcely train with only image 40%, brute-forcely train image with box 62%, suggesting a high location prior in the image, spatial transformer attention crop with gt box 69%.

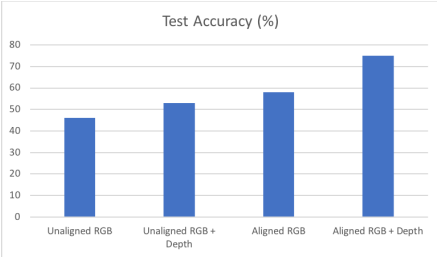


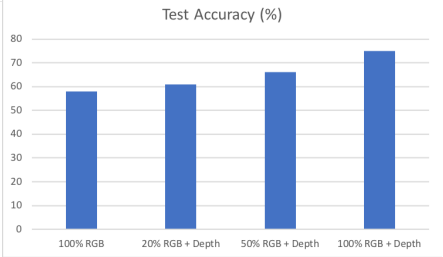
5/3

- Summarize the performance of using the bounding box to align v.s. no aligned, and using RGB + depth v.s. RGB only, on MLT dataset, with 7000 training images and 700 testing images, 6 classes.
- Try to train a residual network to beat the previous VGG network, but residual network obtains worse performance on all different tasks. Will figure out why.
- Start to work on depth based attention model to recognize objects in the cluttered scene (to further study the depth usage). Currently, the baseline without attention is only 40% accuracy on the testing set.





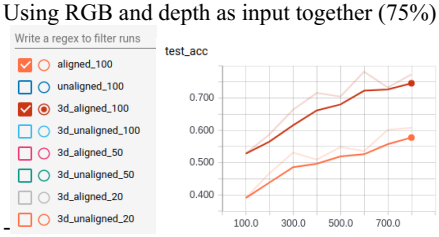




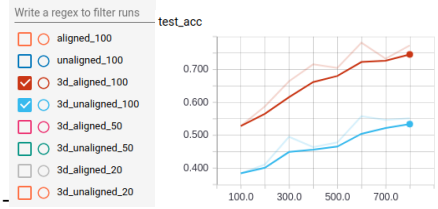
4/26

- Compare the effect of using bounding box to crop an object then classify (crop) v.s. directly recognize object in all possible object scales (wild). The results suggest there is more than 10% (significant) gap on recognizing both 7 classes on VIPER dataset and 6 classes on MLT dataset.
- The simplest random guess baseline on MLT testing is 29% accuracy, without any processing 44% accuracy, with bounding box processing 58% accuracy (16% improvement).

- Using RGB and depth as input together (75%) outperforms using RGB only (58%), significantly (17% improvement).

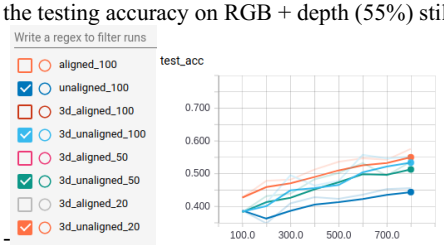


- When objects are not croped with bounding box, using RGB and depth as input together (55%) still outperforms using RGB only (44%), significantly (11% improvement)

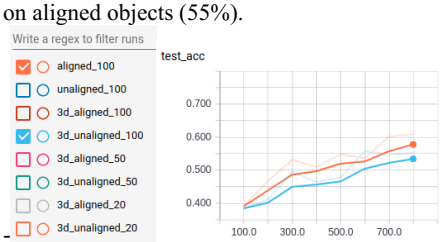


- Even when the number of training data decrease (from 100% to 20%), the testing accuracy on RGB + depth (60%) still outperforms RGB only with 100% training data (58%).

- For unaligned objects, the conclusion is the same, even when the number of training data decrease (from 100% to 20%), the testing accuracy on RGB + depth (55%) still outperforms RGB only with 100% training data (44%)

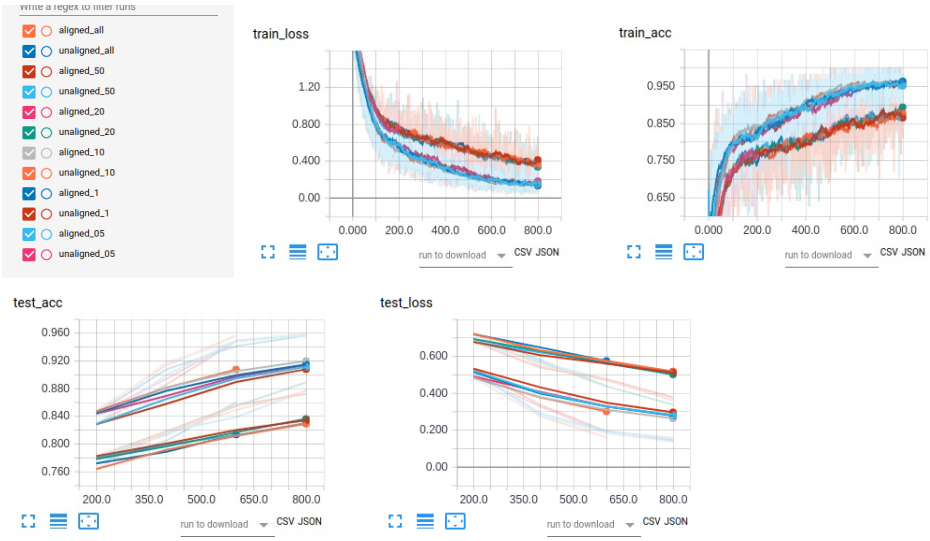


- The accuracy of using RGB + depth on unaligned objects (53%) is approximately equal to the accuracy of using RGB on aligned objects (55%).



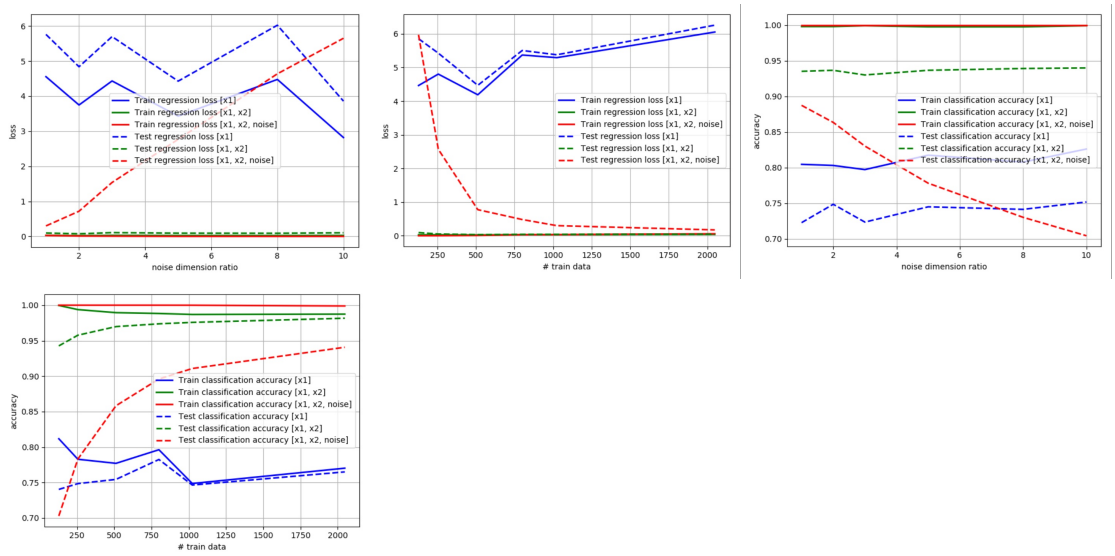
4/5

- Compare the effect of using bounding box to crop an object then classify (crop) v.s. directly recognize object with all possible object scales (wild). The results suggest there is 10% benefit (significant) on recognizing 7 classes in VIPER dataset.
- The baseline is 55% accuracy, without any processing 85% accuracy, with bounding box processing 95%.
- Compare the performance gap between the crop v.s. wild with the amount of training data. It seems there is no significant gap between using more than 30000 images and using only 300 images. Experiments suggest there may be a bug to fix.



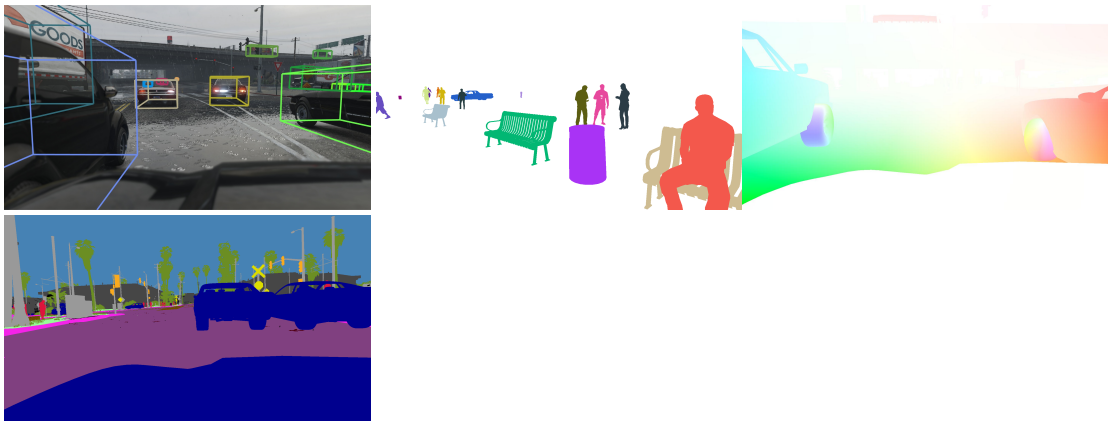
3/29

- Preliminarily study the effect of noise feature on few-shot learning.
- Conclude that if we want the learner to learn fast and accurate with a few training examples, the quality of input feature is significantly important.
- The quality includes two parts: Purity and Integrity (Completeness). Purity means there is no independent noise included in the feature vector, all features are useful. Integrity means the feature is discriminative enough to provide information for classification.
- Derive the equation about bias and variance for the purity and integrity (completeness).
- Conduct experiments on the train / test loss w.r.t. the number of training data and the noise dimensionality ratio (dim(noise) / dim(useful feature))



3/22

- Investigating the reference work on illumination invariance learning, find there are two existing dataset, http://robotics.pme.duth.gr/phos2.html
- Come up a new model for modeling the optical flow on the illumination change. The overall idea is that there will be multiple motions on a single pixel.
- Investigate the new VIPER (playing for benchmarks) dataset, the dataset contains ground truth annotation for videos, including optical flow, camera pose, semantic segmentation, instance segmentation, 3D object detection.
- Will use VIPER dataset to study the lighting change for optical flow estimation and detection for one-shot classification.



3/15

- Writing 2 ECCV papers, one is about 3D object pose estimation, the other is about video hightlight extraction.

3/8

- ECCV writing about 3D object pose estimation with fine-grained objects
- Restart the thought of predictive learning
- Give a talk at group meeting

2/29

- Working on ECCV submission about 3D object pose estimation with fine-grained objects
- Discuss with Yuanpeng about zero-shot Caption-VQA transfer

1/25

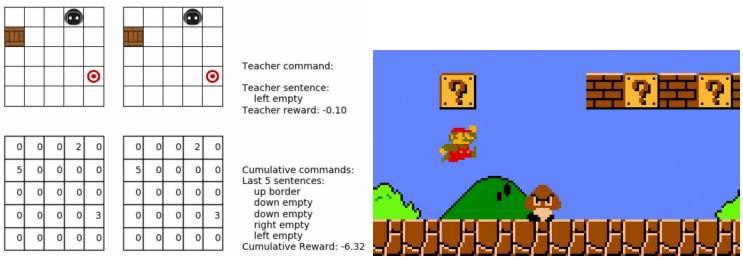
- Read the Book An Introduction to Reinforcement Learning 2 by Richard Sutton.
- Prepare the knowledge for model-based planning.
- No experiment this week.

1/18

- Discuss and prepare an idea for Enrivonment model to improve Exploration

1/11

- Finish building the 2d push box game (preliminarily)
- Train an unsupervised super-mario game with Deep Q Learning, DQN played very poor on this game



1/4

- Start to implement push box game in python based XWorld 2D environment, expected to finish the game development in a week.
- The overall idea is to see whether long-term based planning can help decision making and whether unsupervised learning can help long-term based planning.