SIM-TO-REAL AUTONOMOUS ROBOTIC CONTROL

Zhang-Wei Hong, and Yu-Ming Chen, and Hsuan-kung Yang, and Sih-Ping Lai, and Hsin-Wei Hsiao

Elsa Lab

Department of Computer Science

National Tsing Hua University

OUTLINE

Train in Real World?

Domain Adaption?

HOW?

Depth Map?

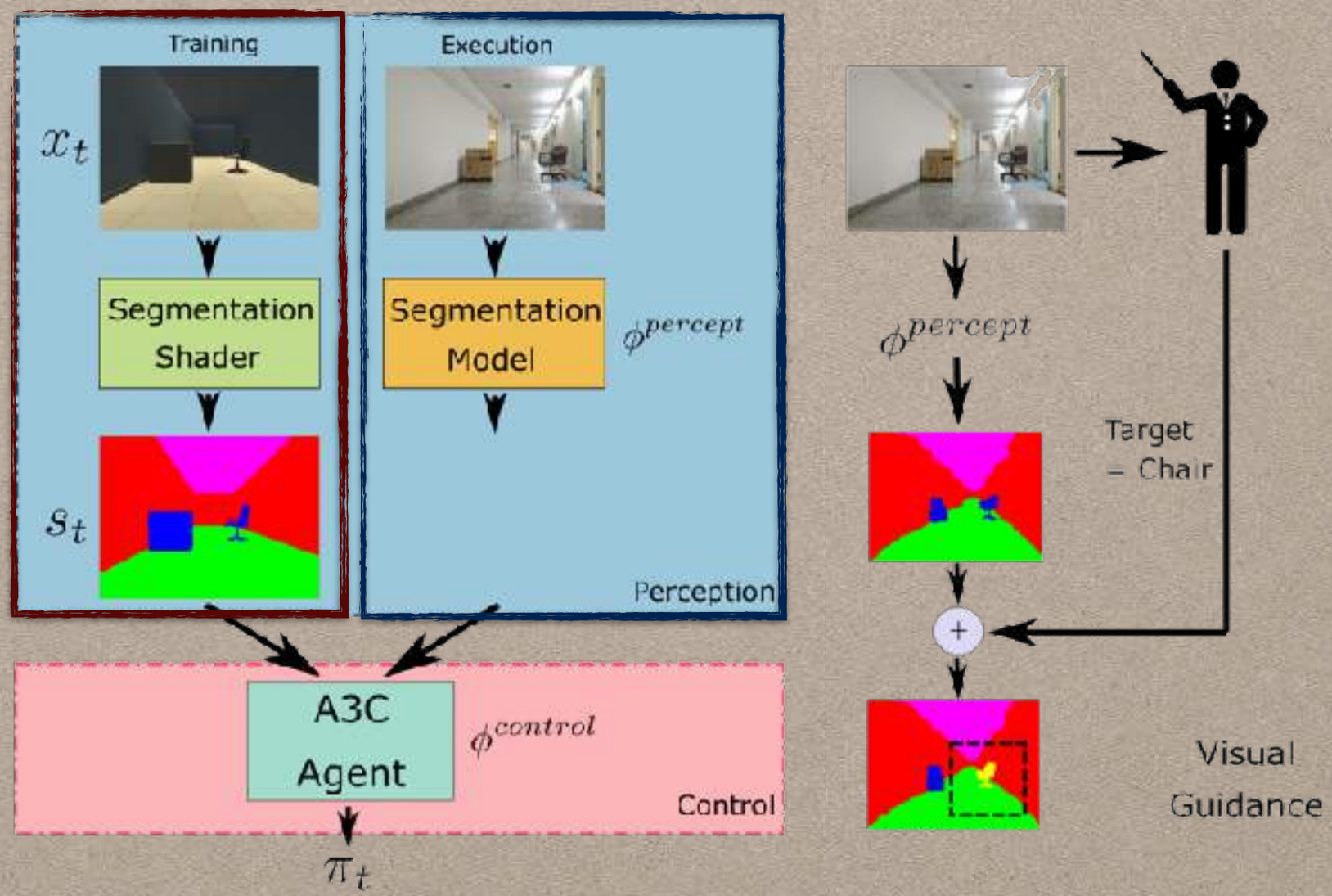
Domain Randomization?

Domain Adaption?

Train in Real World?

BETTER SOLUTION

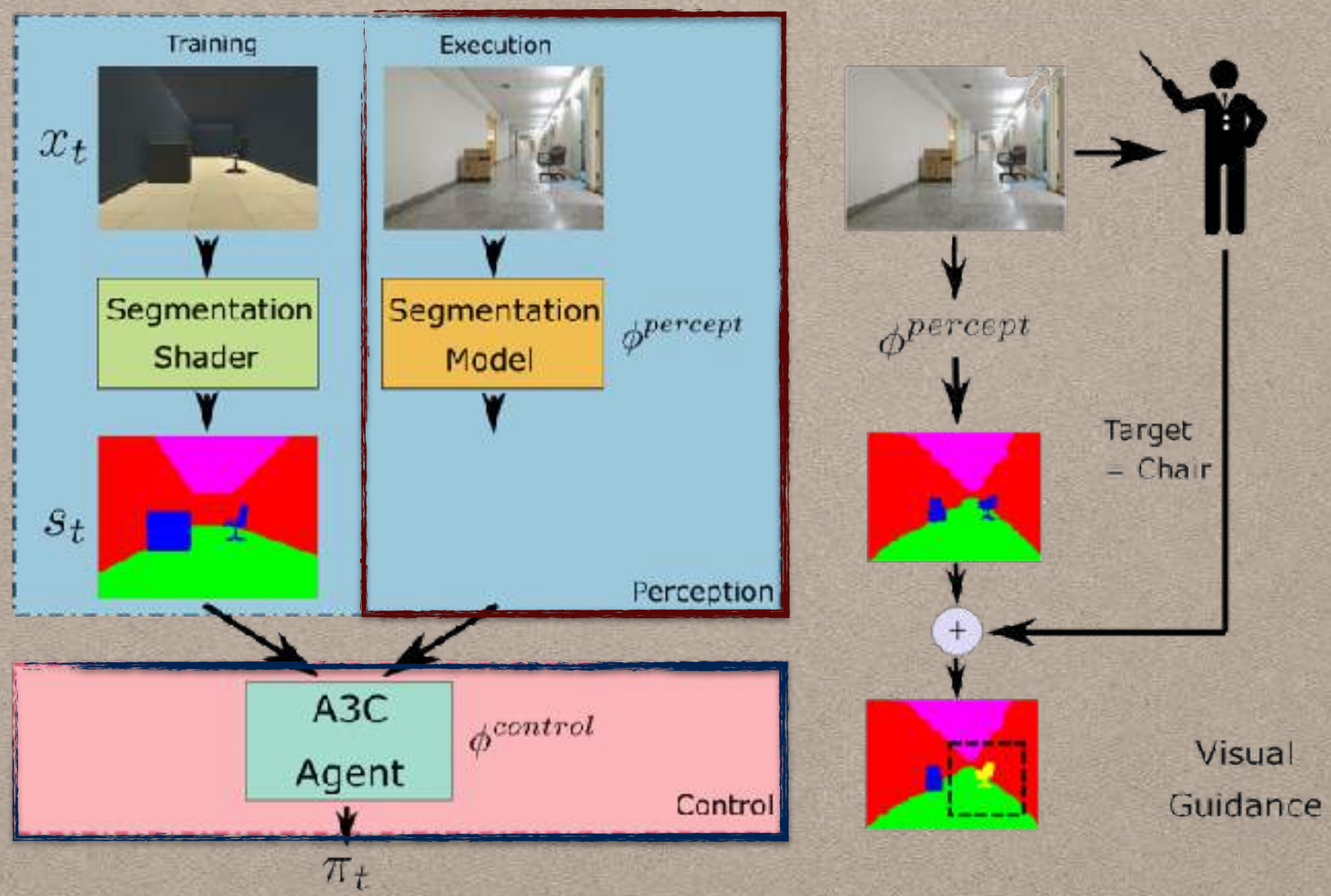
ARCHITECTURE



Train in **simulator**

Apply to **read world**

ARCHITECTURE



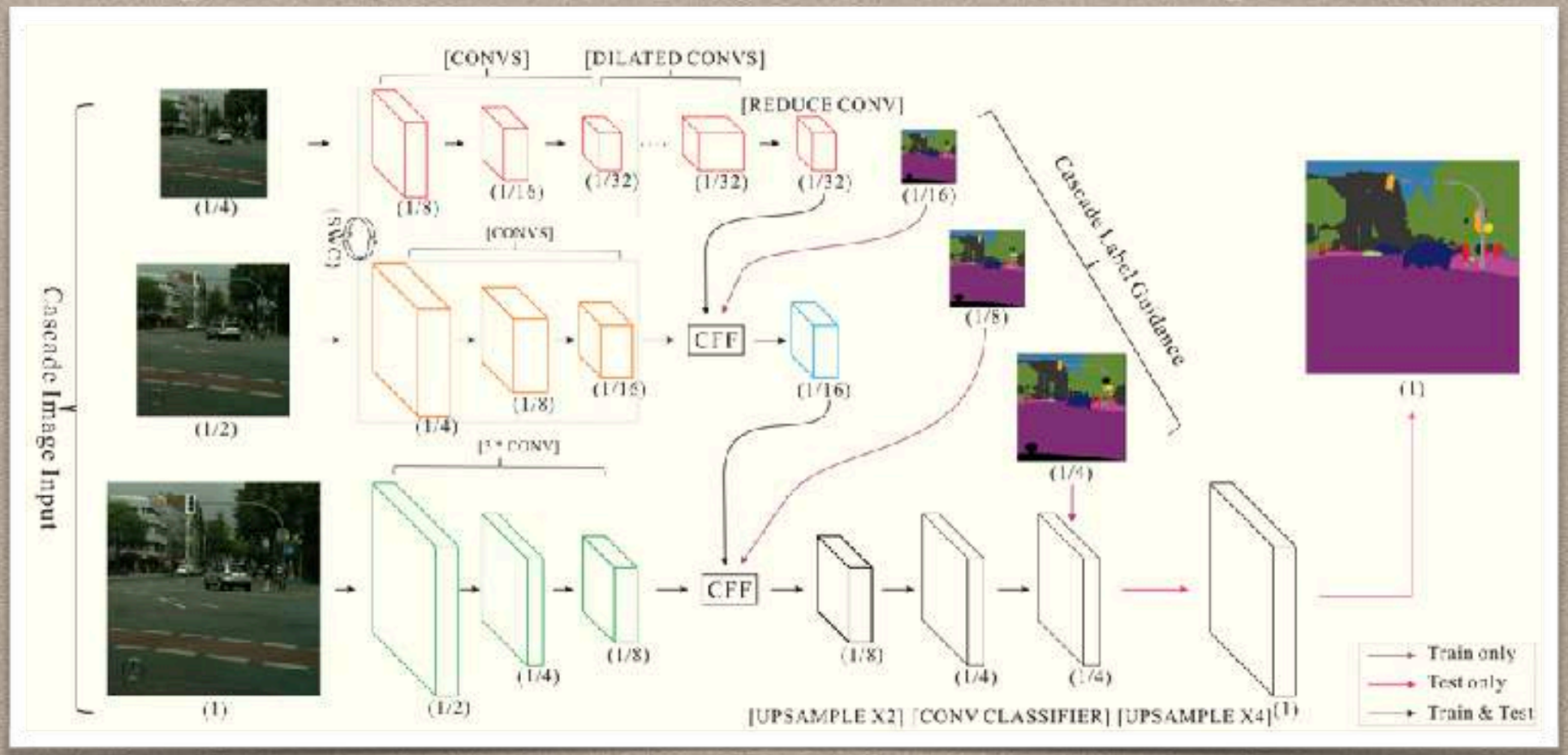
- **Perception module** translates the perceived RGB image to semantic image segmentation
- **Control policy module** performs actions based on translated semantic image segmentation

PERCEPTION MODULE

PERCEPTION MODULE

NETWORK:ICNET[1]

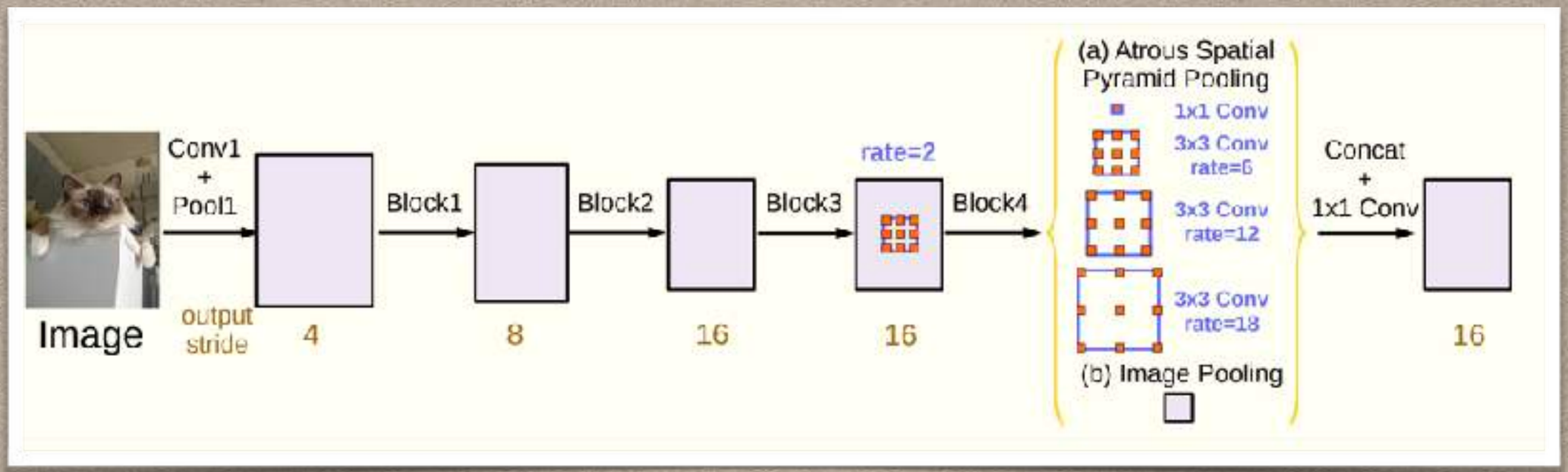
- Incorporates **multi-resolution branches** for accuracy enhancement
- **Real-time inference** on the GPU of NVIDIA Jetson TX2



PERCEPTION MODULE

NETWORK:DEEPLAB[2]

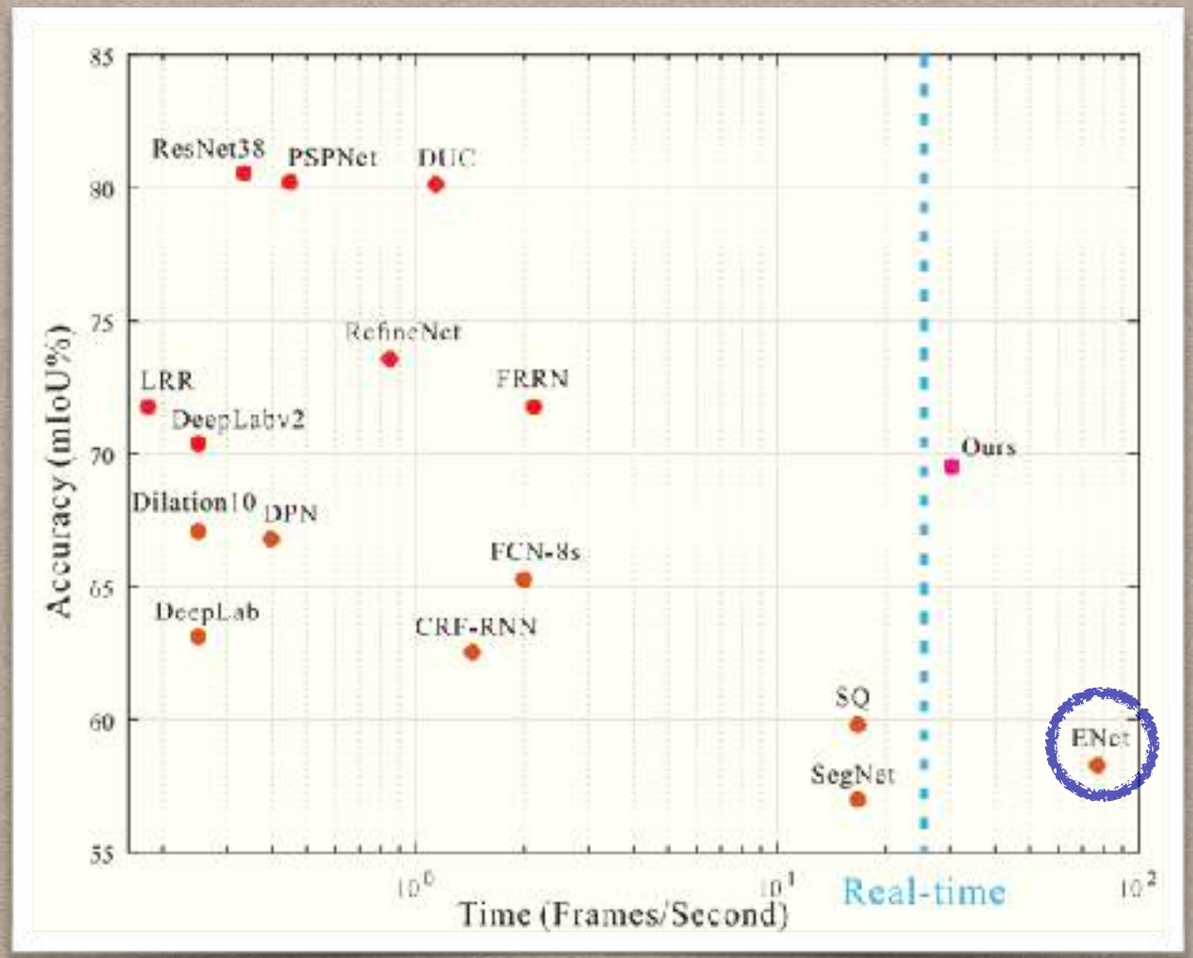
- Using **Atrous convolutions** to **enlarge the field of view** as well as control the resolution of feature
- Using an **Atrous Spatial Pyramid Module(ASPP)** to robustly segment objects at **multiple scales**



PERCEPTION MODULE

NETWORK:ENET[3]

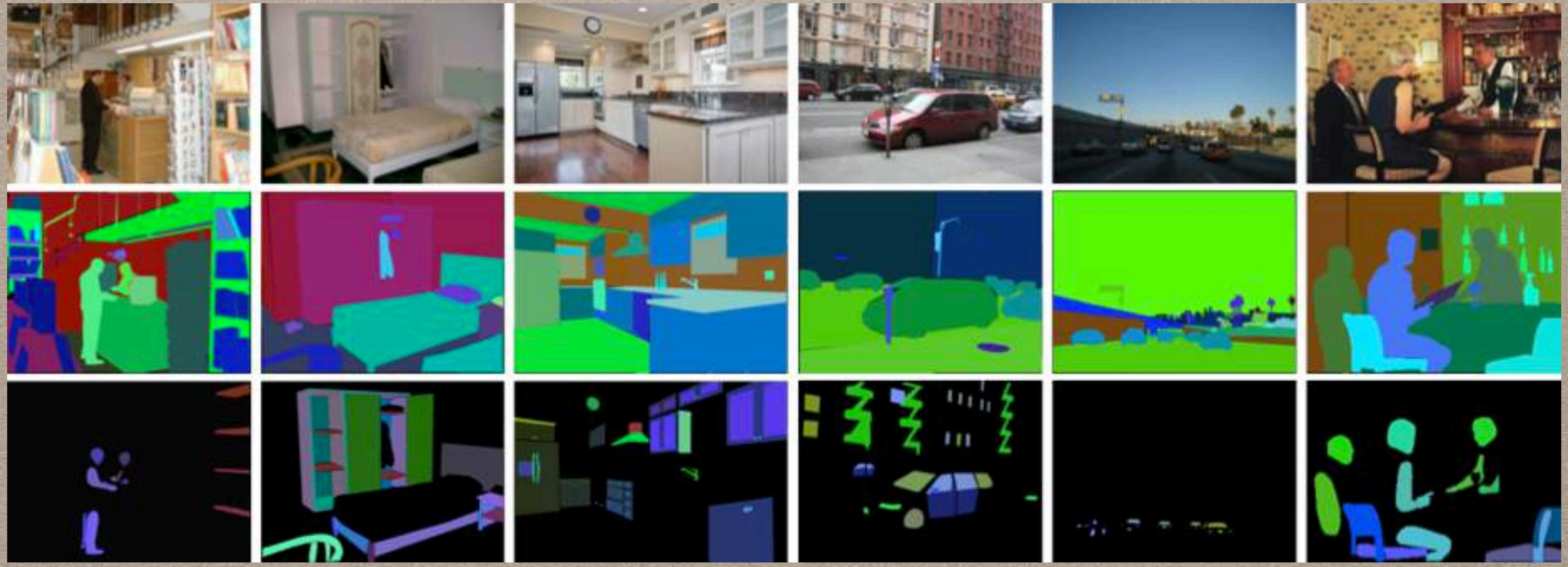
- Sacrifice the accuracy in exchange for the inference speed
- Using less GPU memory to enhance the mobility.



PERCEPTION MODULE

INDOOR DATASET(ADE20K)[4]

- Containing **scene-centric images** annotated with objects
- 20K images for training, 2K images for validation
- Totally **150 semantic categories**



PERCEPTION MODULE

INDOOR DATASET(ADE20K)

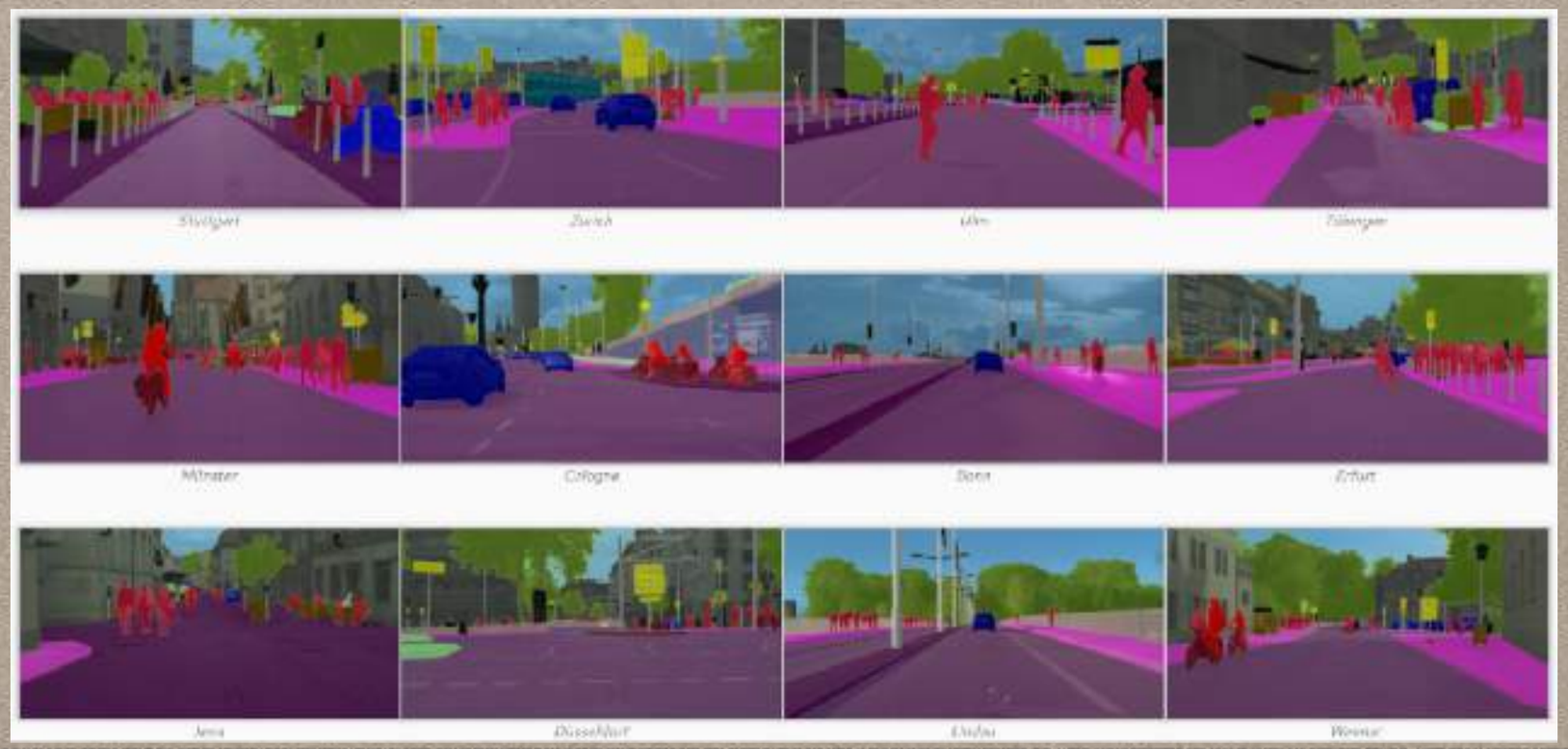
- We **re-labeled the original 150 classes into 27 classes** enhance accuracy and training efficiency
- Re-labeled list :

Original class labels in the dataset	Reduced class labels
window, fence, pillar, door, bulletin board	**wall**
road, ground, field, path, runway	**floor**
bed, cabinet, sofa, table, curtain, chair, shelf, desk	**furniture**
class number larger than 26	**others**

PERCEPTION MODULE

OUTDOOR DATASET(CITYSCAPE)[5]

- Contains a diverse set of video sequences **recorded in street scenes from 50 different cities**
- 5K annotated images.
- Totally **19 classes**

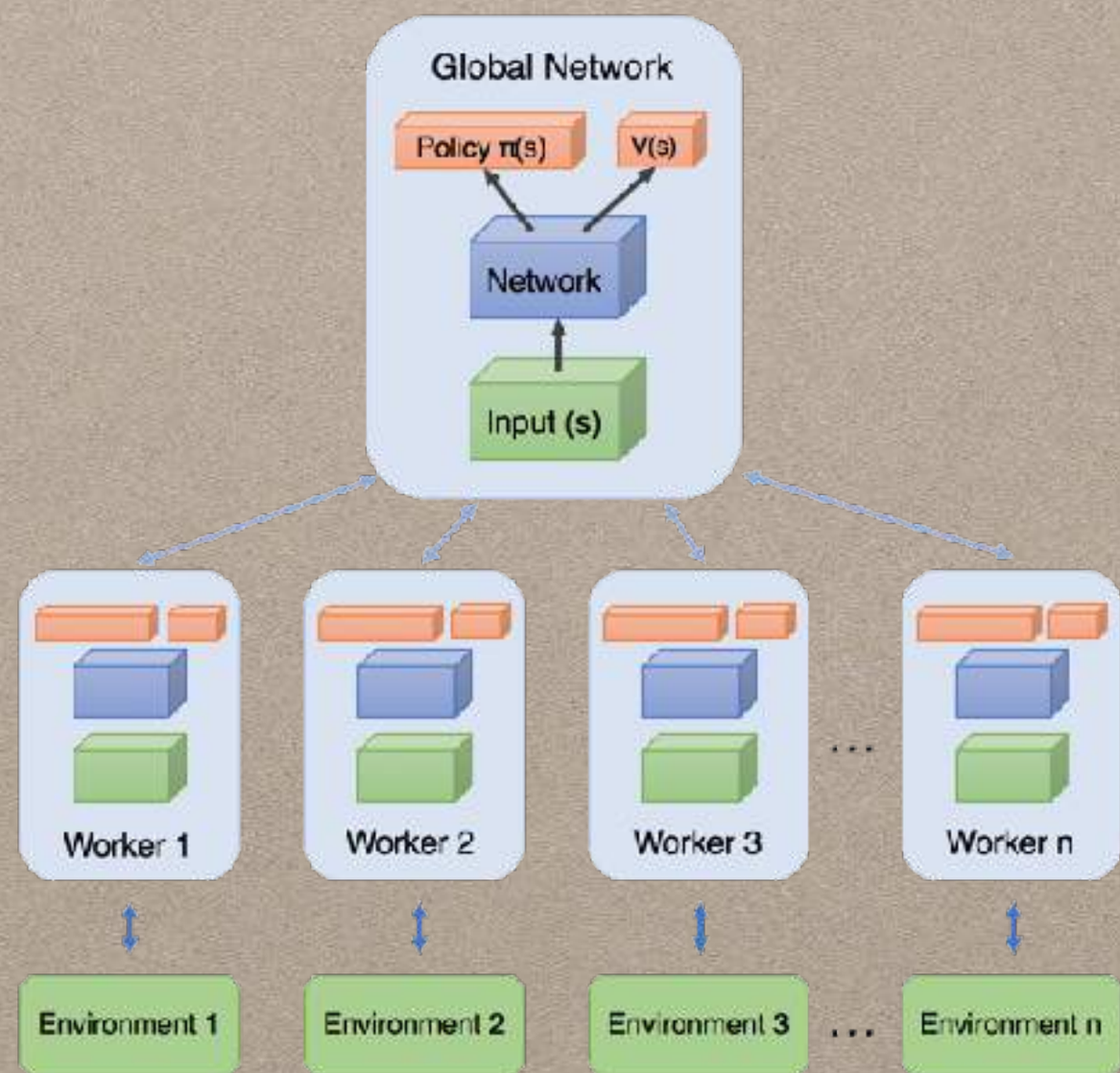


CONTROL POLICY MODULE

CONTROL POLICY

NETWORK:A3C [6]

- Parallelized training for reinforcement learning agents
- No replay buffer



CONTROL POLICY

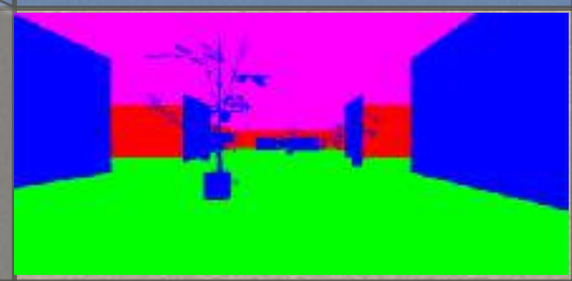
BASELINE MODULE

- Seg : Segmentation
- DR : Domain Randomization
- S : Stacked

Model	Input dimension	Input format
Seg (Ours)	84 x 84 x 3	RGB Frame
Seg-S (Ours)	84 x 84 x 3 x 4	RGB Frame
DR-A3C	84 x 84 x 3	RGB Frame
DR-A3C-S	84 x 84 x 3 x 4	RGB Frame
ResNet-A3C	224 x 224 x 3	RGB Frame
Depth-A3C	84 x 84 x 1	Depth Map
Depth-A3C-S	84 x 84 x 4	Depth Map

CONTROL POLICY

TRAINING SCENE

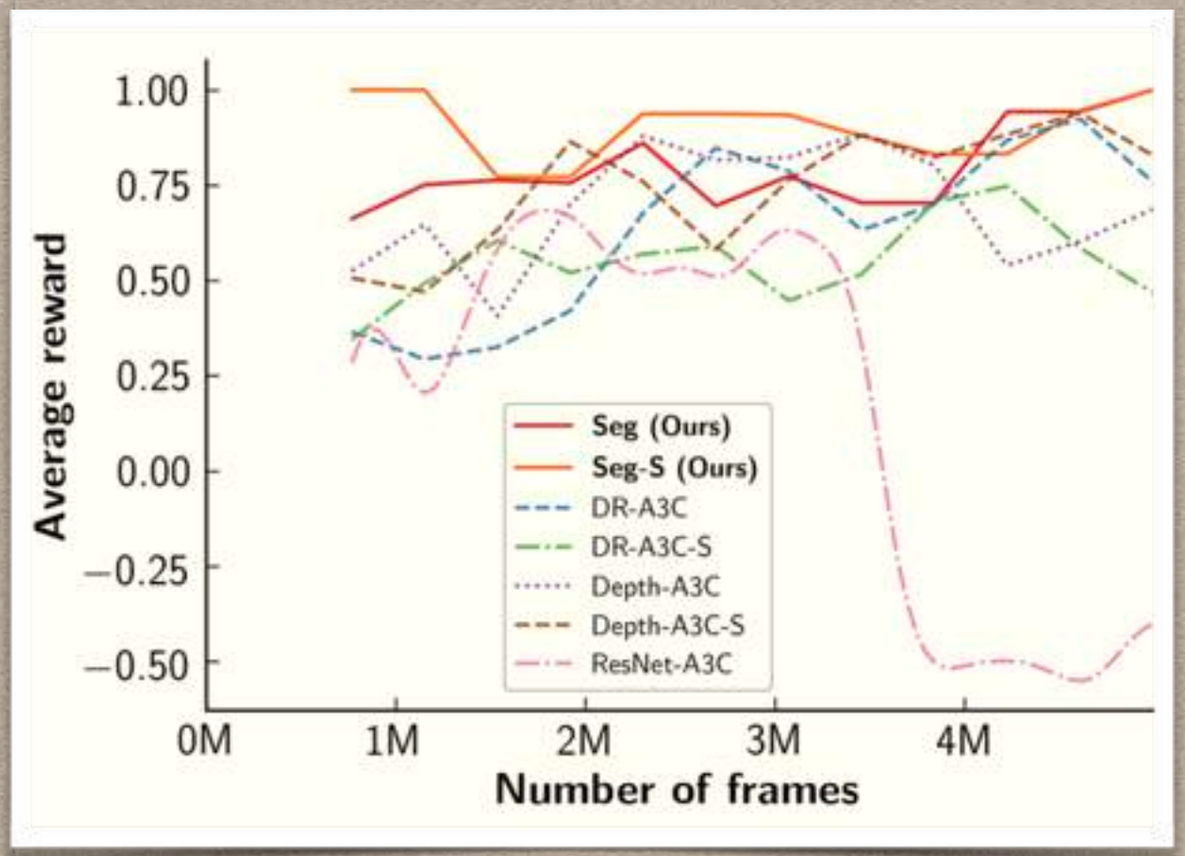
	Segmentation	Domain Randomization	Depth Perception
Navigation			
Target Following			

EXPERIMENTAL RESULTS

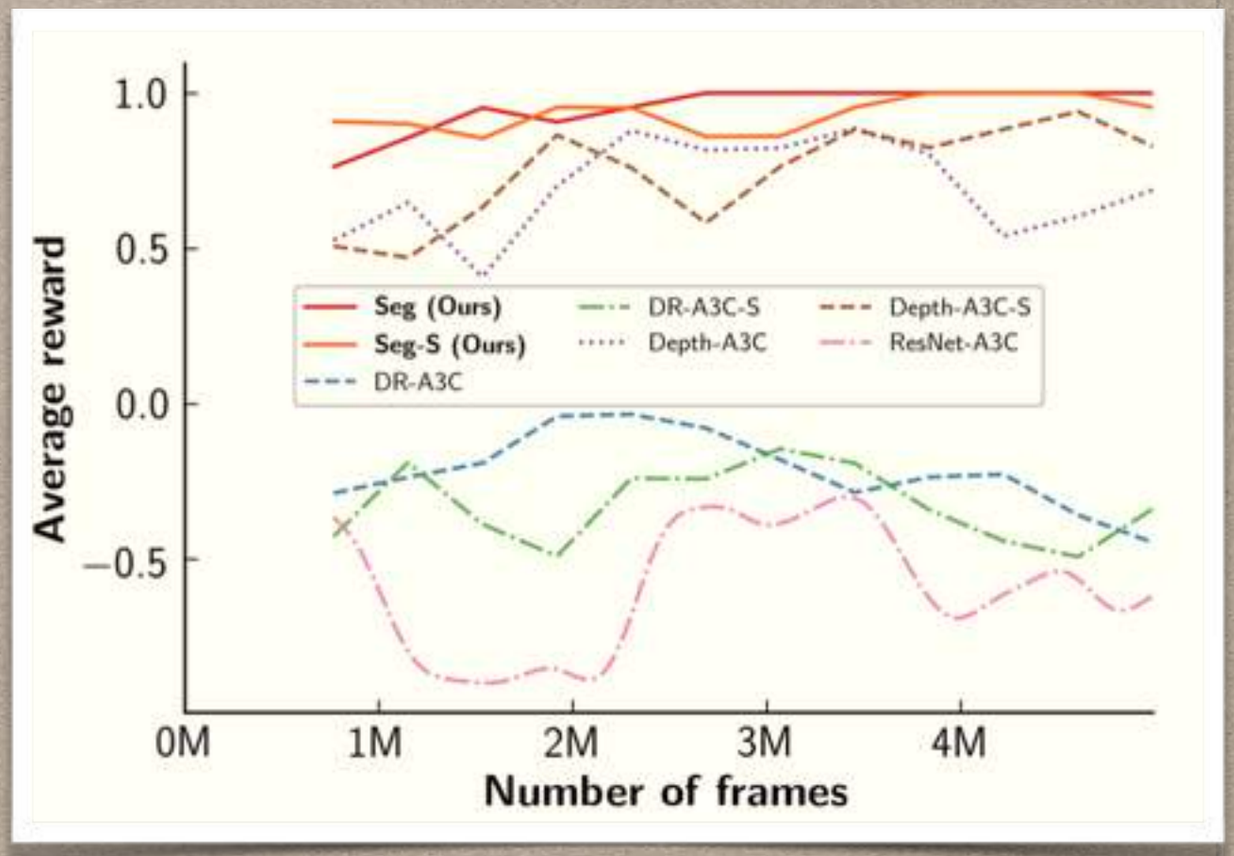
EXPERIMENTAL RESULTS

LEARNING CURVES

- Average rewards of 10 iterations after each training episode



Obstacle Avoidance

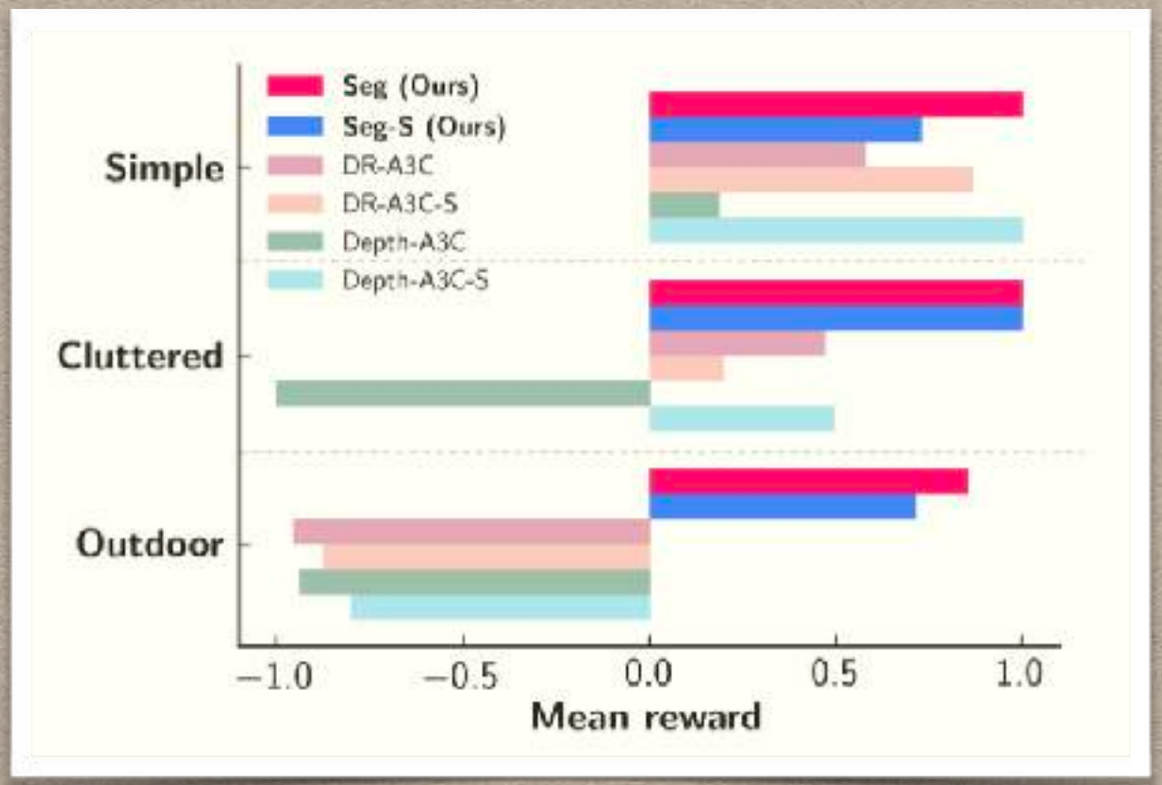


Target Following

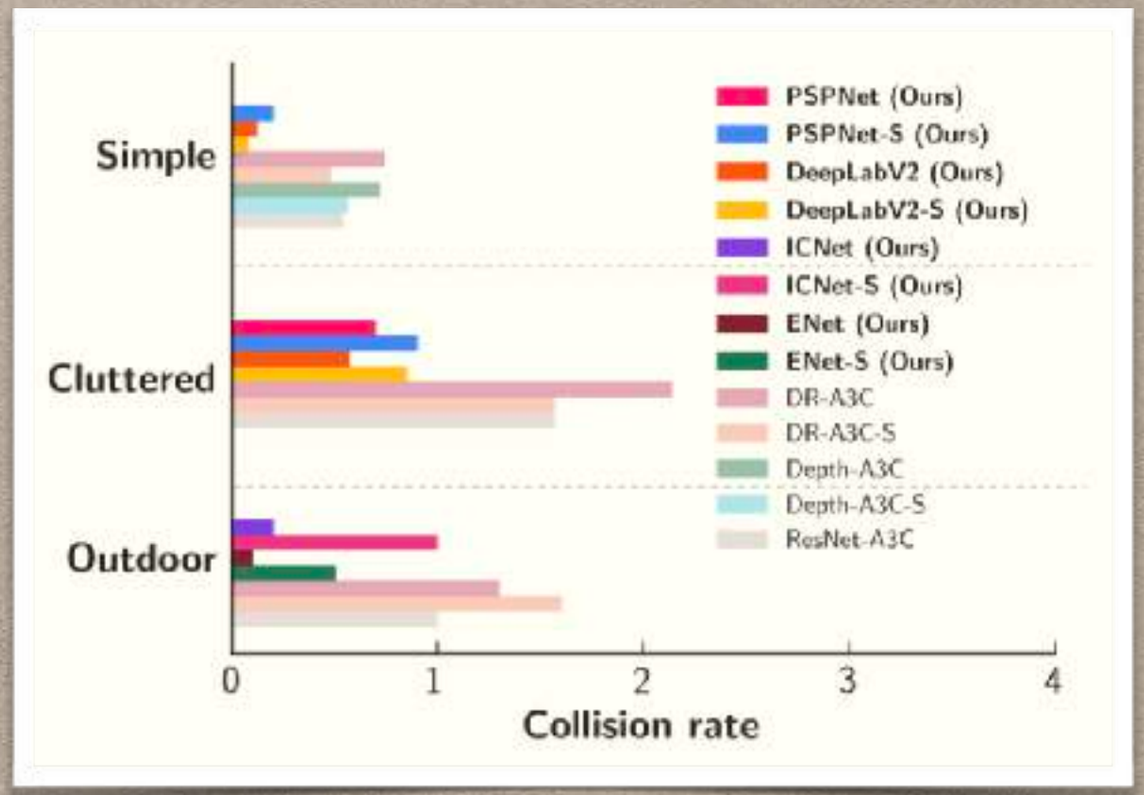
EXPERIMENTAL RESULTS

EVALUATION:OBSTACLE AVOIDANCE

- Mean rewards of 100 episodes in the simulated environment
- Average collision rates within 1 minute in the real world



Simulated Environment

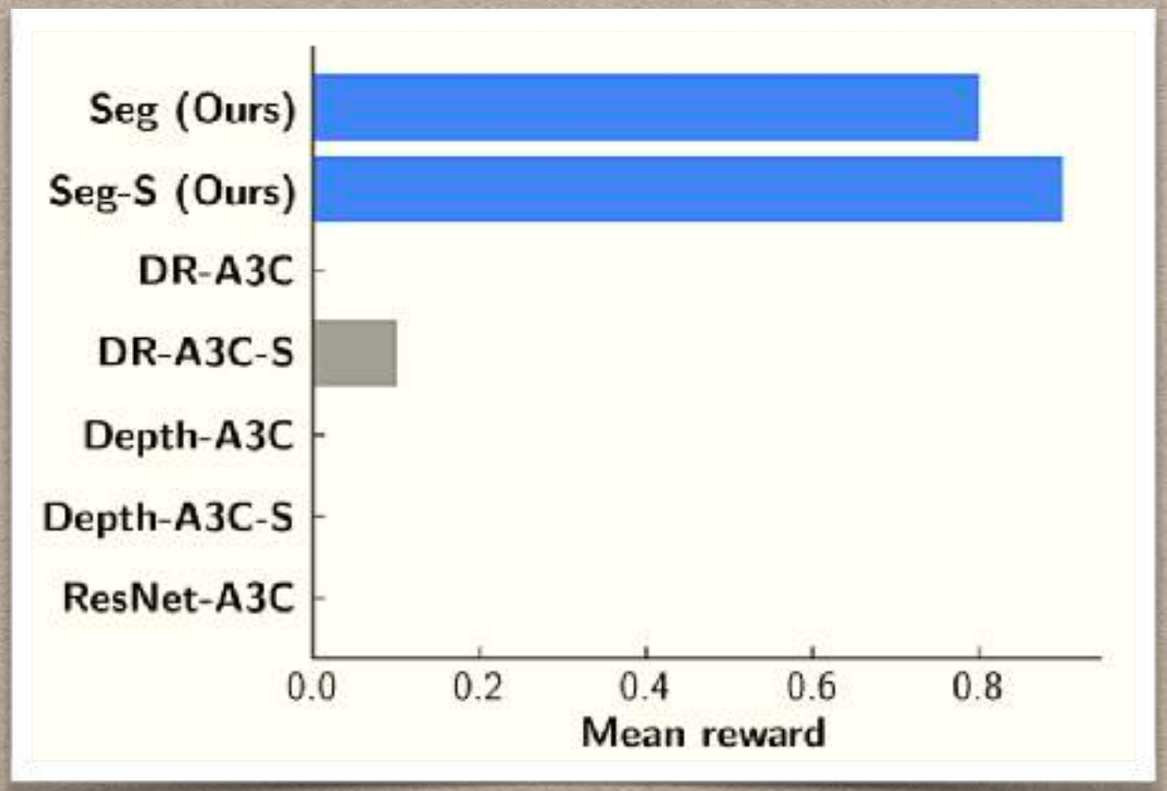


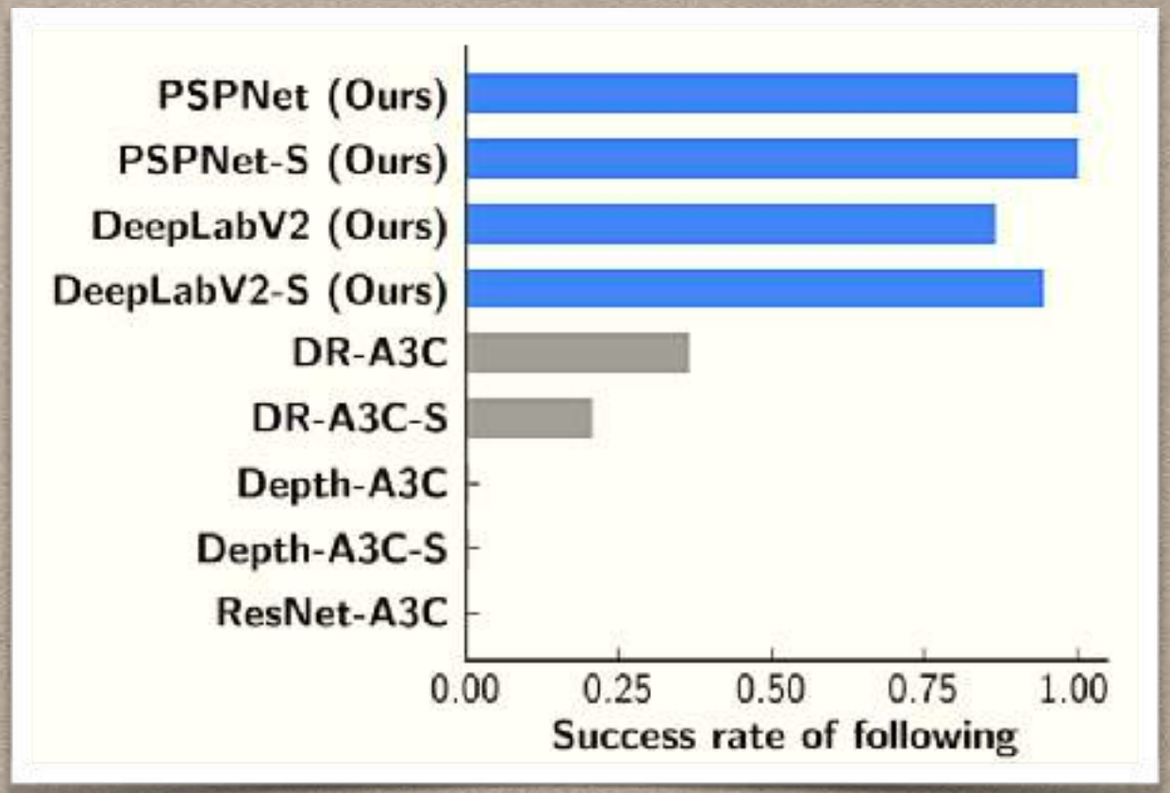
Real World

EXPERIMENTAL RESULTS

EVALUATION:TARGET FOLLOWING

- Mean rewards of 100 episodes in the simulated environment
- Average success rates in the real world





Simulated Environment

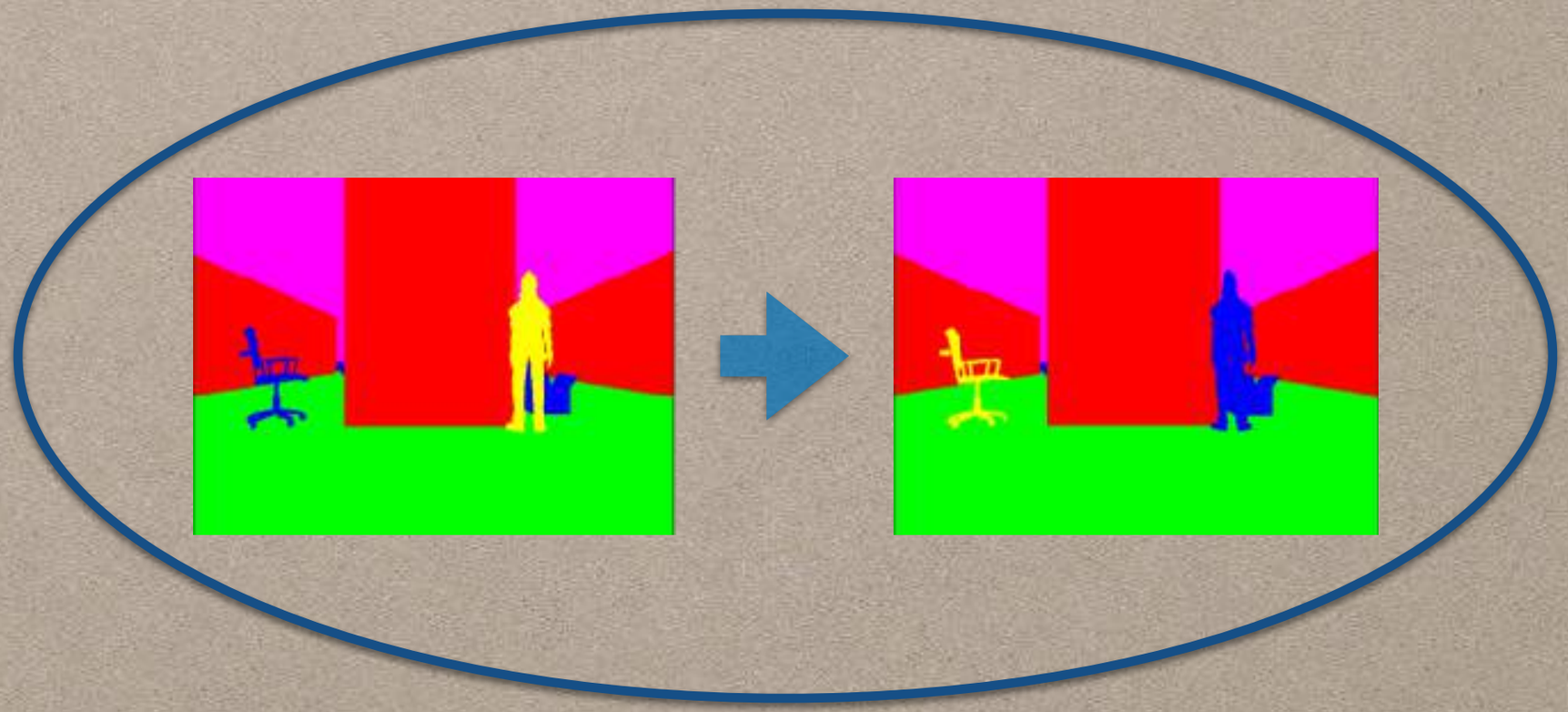
Real World

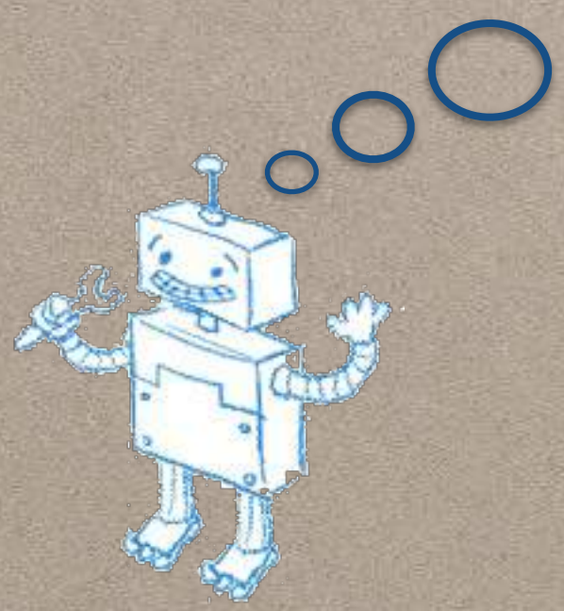
VISUAL GUIDANCE

VISUAL GUIDANCE

SWITCHING-TARGET FOLLOWING TASK

- The visual guidance module can also alter a target following robot's objective online by modifying the target label to a new one.





VISUAL GUIDANCE

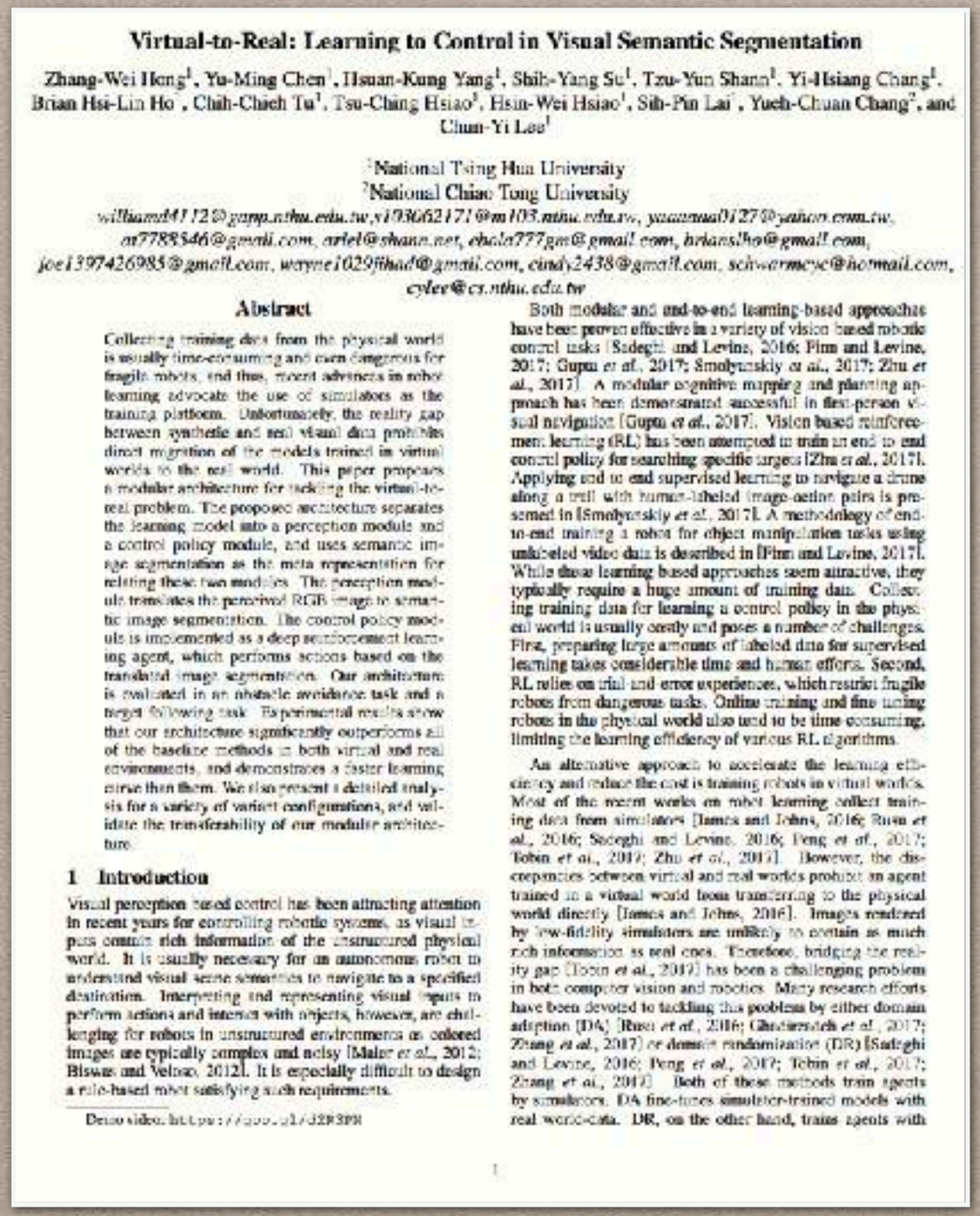
SWITCHING-TARGET FOLLOWING TASK

- The use of scene semantics as the meta-state gives the proposed architecture extra flexibility
- Our modular architecture allows a visual-guidance module to be augmented to perform even more complex tasks by manipulating the meta-state
- Visual guidance does not require any retraining, fine-tuning, or extra data

	Virtual World (Mean Reward)	Real World (Success Rate)
Seg	0.824	80%
Seg-S	0.925	90%

CONTRIBUTIONS

CONTRIBUTIONS



Virtual-to-Real: Learning to Control in Visual Semantic Segmentation

Zhang-Wei Hong[1], Yu-Ming Chen[1], Hsuan-Kung Yang[1], Shih-Yang Su[1], Tzu-Yun Shann[1], Yi-Hsiang Chang[1], Brian Hsi-Lin Ho[1], Chih-Chieh Tu[1], Tsu-Ching Hsiao[1], Hsin-Wei Hsiao[1], Sih-Pin Lai[1], Yueh-Chuan Chang[2], and Chun-Yi Lee[1]

[1]National Tsing Hua University
[2]National Chiao Tong University
williamd4112@gapp.nthu.edu.tw, s1936621710@m103.nthu.edu.tw, yuanmai01270@yahoo.com.tw, ar7788546@gmail.com, arlet@shann.net, ehola777gm@gmail.com, briansl0o@gmail.com, joe1397426985@gmail.com, wayne1029fihad@gmail.com, cindy2438@gmail.com, schwarmeye@hotmail.com, cylee@cs.nthu.edu.tw

Abstract

Collecting training data from the physical world is usually time-consuming and even dangerous for fragile robots, and thus, recent advances in robot learning advocate the use of simulators as the training platform. Unfortunately, the reality gap between synthetic and real visual data prohibits direct migration of the models trained in virtual worlds to the real world. This paper proposes a modular architecture for tackling the virtual-to-real problem. The proposed architecture separates the learning model into a perception module and a control policy module, and uses semantic image segmentation as the meta representation for relating these two modules. The perception module translates the perceived RGB image to semantic image segmentation. The control policy module is implemented as a deep reinforcement learning agent, which performs actions based on the translated image segmentation. Our architecture is evaluated in an obstacle avoidance task and a target following task. Experimental results show that our architecture significantly outperforms all of the baseline methods in both virtual and real environments, and demonstrates a faster learning curve than them. We also present a detailed analysis for a variety of variant configurations, and validate the transferability of our modular architecture.

1 Introduction

Visual perception based control has been attracting attention in recent years for controlling robotic systems, as visual inputs contain rich information of the unstructured physical world. It is usually necessary for an autonomous robot to understand visual scene semantics to navigate to a specified destination. Interpreting and representing visual inputs to perform actions and interact with objects, however, are challenging for robots in unstructured environments as colored images are typically complex and noisy [Maire et al., 2012; Biswas and Veloso, 2012]. It is especially difficult to design a rule-based robot satisfying such requirements.

Demo video: https://youtu.be/fJ2B3PR

Both modular and end-to-end learning-based approaches have been proven effective in a variety of vision based robotic control tasks [Sadeghi and Levine, 2016; Finn and Levine, 2017; Gupta et al., 2017; Smolyanskiy et al., 2017; Zhu et al., 2017]. A modular cognitive mapping and planning approach has been demonstrated successful in first-person visual navigation [Gupta et al., 2017]. Vision based reinforcement learning (RL) has been attempted to train an end to end control policy for searching specific targets [Zhu et al., 2017]. Applying end to end supervised learning to navigate a drone along a trail with human-labeled image-action pairs is presented in [Smolyanskiy et al., 2017]. A methodology of end-to-end training a robot for object manipulation tasks using unlabeled video data is described in [Finn and Levine, 2017]. While these learning based approaches seem attractive, they typically require a huge amount of training data. Collecting training data for learning a control policy in the physical world is usually costly and poses a number of challenges. First, preparing large amounts of labeled data for supervised learning takes considerable time and human efforts. Second, RL relies on trial-and error experiences, which restrict fragile robots from dangerous tasks. Online training and fine tuning robots in the physical world also tend to be time consuming, limiting the learning efficiency of various RL algorithms.

An alternative approach to accelerate the learning efficiency and reduce the cost is training robots in virtual worlds. Most of the recent works on robot learning collect training data from simulators [James and Johns, 2016; Rusu et al., 2016; Sadeghi and Levine, 2016; Peng et al., 2017; Tobin et al., 2017; Zhu et al., 2017]. However, the discrepancies between virtual and real worlds prohibit an agent trained in a virtual world from transferring to the physical world directly [James and Johns, 2016]. Images rendered by low-fidelity simulators are unlikely to contain as much rich information as real ones. Therefore, bridging the reality gap [Tobin et al., 2017] has been a challenging problem in both computer vision and robotics. Many research efforts have been devoted to tackling this problem by either domain adaption (DA) [Rusu et al., 2016; Gheshlaghi et al., 2017; Zhang et al., 2017] or domain randomization (DR) [Sadeghi and Levine, 2016; Peng et al., 2017; Tobin et al., 2017; Zhang et al., 2017]. Both of these methods train agents by simulators. DA fine-tunes simulator-trained models with real world-data. DR, on the other hand, trains agents with

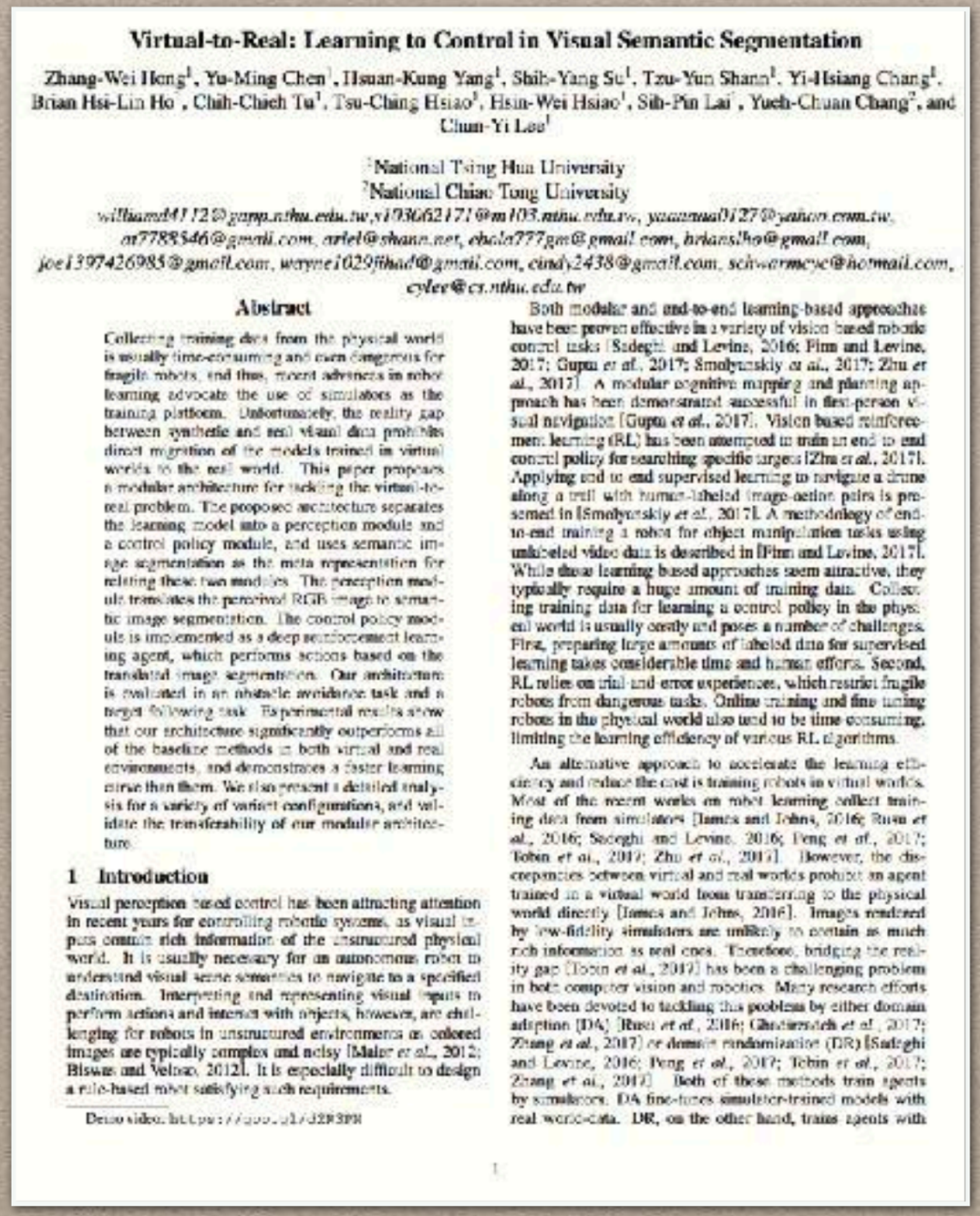
1

https://arxiv.org/abs/1802.00285

- A new modular learning-based architecture which separates the vision-based robotic learning model into a perception module and a control policy module.
- A novel concept for bridging the reality gap via the use of semantic image segmentation
- A simple methodology for directly transferring the control policy learned in virtual environments to the real world.

CONTRIBUTIONS

CONTINUED



Virtual-to-Real: Learning to Control in Visual Semantic Segmentation

Zhang-Wei Hong[1], Yu-Ming Chen[1], Hsuan-Kung Yang[1], Shih-Yang Su[1], Tzu-Yun Shann[1], Yi-Hsiang Chang[1], Brian Hsi-Lin Ho[1], Chih-Chieh Tu[1], Tsu-Ching Hsiao[1], Hsin-Wei Hsiao[1], Sih-Pin Lai[1], Yueh-Chuan Chang[2], and Chun-Yi Lee[1]

[1]National Tsing Hua University
[2]National Chiao Tong University
williamd4112@gapp.nthu.edu.tw, s1936621710@m103.nthu.edu.tw, yuanmai01270@yahoo.com.tw, ar7788546@gmail.com, arlet@shann.net, ehola777gm@gmail.com, brianslho@gmail.com, joe1397426985@gmail.com, wayne1029fihad@gmail.com, cindy2438@gmail.com, schwarmeye@hotmail.com, cylee@cs.nthu.edu.tw

Abstract

Collecting training data from the physical world is usually time-consuming and even dangerous for fragile robots, and thus, recent advances in robot learning advocate the use of simulators as the training platform. Unfortunately, the reality gap between synthetic and real visual data prohibits direct migration of the models trained in virtual worlds to the real world. This paper proposes a modular architecture for tackling the virtual-to-real problem. The proposed architecture separates the learning model into a perception module and a control policy module, and uses semantic image segmentation as the meta representation for relating these two modules. The perception module translates the perceived RGB image to semantic image segmentation. The control policy module is implemented as a deep reinforcement learning agent, which performs actions based on the translated image segmentation. Our architecture is evaluated in an obstacle avoidance task and a target following task. Experimental results show that our architecture significantly outperforms all of the baseline methods in both virtual and real environments, and demonstrates a faster learning curve than them. We also present a detailed analysis for a variety of variant configurations, and validate the transferability of our modular architecture.

1 Introduction

Visual perception based control has been attracting attention in recent years for controlling robotic systems, as visual inputs contain rich information of the unstructured physical world. It is usually necessary for an autonomous robot to understand visual scene semantics to navigate to a specified destination. Interpreting and representing visual inputs to perform actions and interact with objects, however, are challenging for robots in unstructured environments as colored images are typically complex and noisy [Maier et al., 2012; Biswas and Veloso, 2012]. It is especially difficult to design a rule-based robot satisfying such requirements.

Demo video: https://youtu.be/fd2E3PR

Both modular and end-to-end learning-based approaches have been proven effective in a variety of vision based robotic control tasks [Sadeghi and Levine, 2016; Finn and Levine, 2017; Gupta et al., 2017; Smolyanskiy et al., 2017; Zhu et al., 2017]. A modular cognitive mapping and planning approach has been demonstrated successful in first-person visual navigation [Gupta et al., 2017]. Vision based reinforcement learning (RL) has been attempted to train an end to end control policy for searching specific targets [Zhu et al., 2017]. Applying end to end supervised learning to navigate a drone along a trail with human-labeled image-action pairs is presented in [Smolyanskiy et al., 2017]. A methodology of end-to-end training a robot for object manipulation tasks using unlabeled video data is described in [Finn and Levine, 2017]. While these learning based approaches seem attractive, they typically require a huge amount of training data. Collecting training data for learning a control policy in the physical world is usually costly and poses a number of challenges. First, preparing large amounts of labeled data for supervised learning takes considerable time and human efforts. Second, RL relies on trial-and error experiences, which restrict fragile robots from dangerous tasks. Online training and fine tuning robots in the physical world also tend to be time consuming, limiting the learning efficiency of various RL algorithms.

An alternative approach to accelerate the learning efficiency and reduce the cost is training robots in virtual worlds. Most of the recent works on robot learning collect training data from simulators [James and Johns, 2016; Rusu et al., 2016; Sadeghi and Levine, 2016; Peng et al., 2017; Tobin et al., 2017; Zhu et al., 2017]. However, the discrepancies between virtual and real worlds prohibit an agent trained in a virtual world from transferring to the physical world directly [James and Johns, 2016]. Images rendered by low-fidelity simulators are unlikely to contain as much rich information as real ones. Therefore, bridging the reality gap [Tobin et al., 2017] has been a challenging problem in both computer vision and robotics. Many research efforts have been devoted to tackling this problem by either domain adaption (DA) [Rusu et al., 2016; Gharbiatalab et al., 2017; Zhang et al., 2017] or domain randomization (DR) [Sadeghi and Levine, 2016; Peng et al., 2017; Tobin et al., 2017; Zhang et al., 2017]. Both of these methods train agents by simulators. DA fine-tunes simulator-trained models with real world-data. DR, on the other hand, trains agents with

1

- A way to migrate the operational environment of a robot without further fine-tuning its control policy module.
- A visual-guidance module for altering the behavior of the robot via adjusting the meta-state representations.

https://arxiv.org/abs/1802.00285

THANK YOU FOR YOUR ATTENTION

REFERENCE

[1] Y. You, X. Pan, Z. Wang, and C. Lu. Vir- tual to real reinforcement learning for autonomous driving. *arXiv:1704.03952*, Sep. 2017.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Analysis and Ma- chine Intelligence*, Jun. 2016.

[3] A. Paszke, A. Chaurasia, S. Kim, and E. Cu- lurciello. ENet: A deep neural network architecture for real-time semantic segmentation. *arXiv:1606.02147*, 2016.

[4] Y. You, X. Pan, Z. Wang, and C. Lu. Vir- tual to real reinforcement learning for autonomous driving. *arXiv:1704.03952*, Sep. 2017.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3213-3223, Jun. 2016. Cityscape

[6] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proc. Int. Conf. Machine Learning (ICML)*, pp. 1928-1937, Jul. 2016.