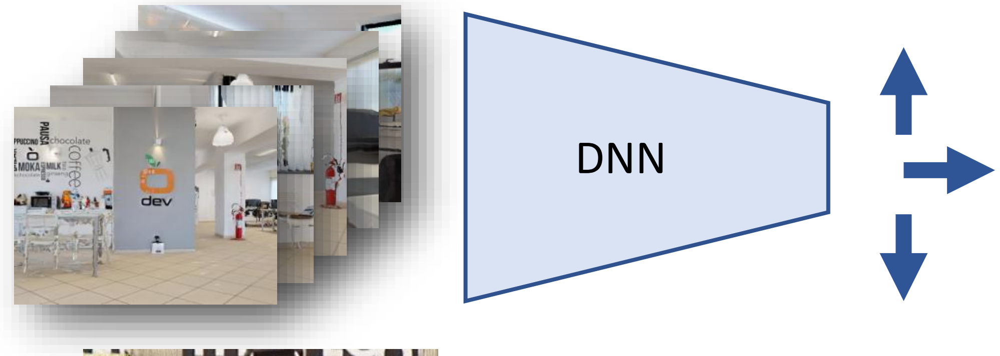
On Embodied Visual Navigation in Real Environments Through Habitat

Marco Rosano[1,3], Antonino Furnari[1], Luigi Gulino[3], Giovanni Maria Farinella[1,2]

[1]FPV@IPLAB - Department of Mathematics and Computer Science, University of Catania, Italy
[2]Cognitive Robotics and Social Sensing Laboratory, ICAR-CNR, Palermo, Italy
[3]OrangeDev s.r.l., Firenze, Italy
marco.rosano@unict.it, furnari@dmi.unict.it, luigi.gulino@orangedev.it, gfarinella@dmi.unict.it

UNIVERSITÀ degli STUDI di CATANIA

Deep Learning Model for Navigation



- Recent Deep Learning approaches have shown that is possible to learn navigation policies in a end-toend fashion from images

- Major drawback: they require to collect a lot of experience, acting and interactiong with the enviroment

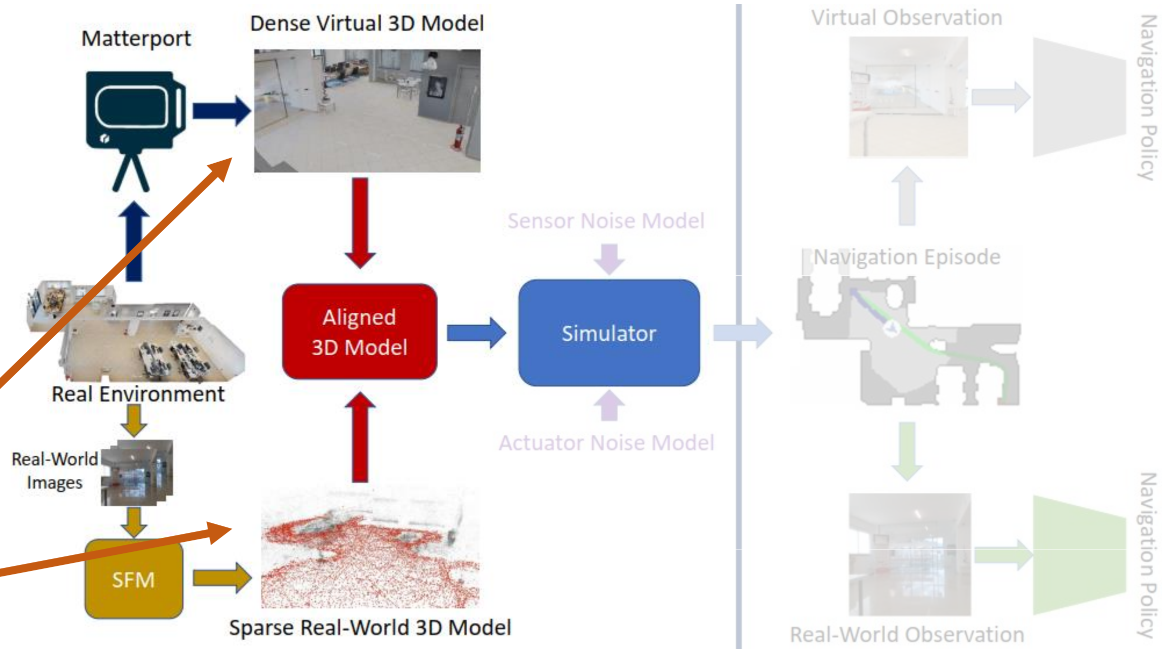
- Collecting the required experience in the real word is difficult
  - Robots are costly and fragile
  - Perform the navigation episodes requires time

- Possible solution: collect the experience in simulation
  - Pros: efficient, scalable
  - Cons: the learned policies perform poorly when deployed in the real world
    - Differences in the appearance
    - No sensor and actuator noise
    - Simplified physical interaction

Train in Simulation with Real-World Images

- We propose a tool based on the popular Habitat simulator[1] to train and evaluate visual navigation policies:
  - Entirely in simulation, avoiding the deploy on a physical robot
  - Using virtual and real observations
  - With realistic sensor and actuator noise

- Acquisition of two 3D models of the same environment:
  - Virtual 3D model, geometrically accurate but with a limited photorealism
  - Real-World 3D model, geometrically inaccurate but containing photo-realistic images

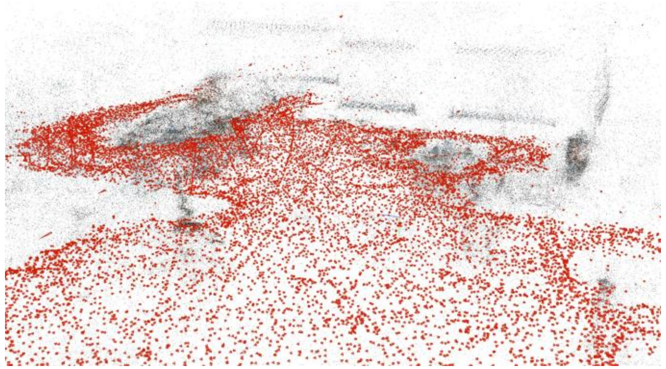


[1] Savva et al., «Habitat: A platform for embodied ai research». In ICCV 2019

3D Models Construction and Alignment

- ~25k Real-world Images
- Collected with a robot following a simple exploration policy

Structure-from-Motion
Reconstruction Pipeline



Sparse Real-World 3D Model

- Matterport 3D Scanner

Dense Virtual 3D Model

- Additional 6k virtual images, with minimum sampling costs
- Used to align the real-world model to the virtual one

Image-Based
Alignment Procedure

Aligned 3D Models

Rosano M., Furnari A., Gulino L., Farinella G.M. *«On Embodied Visual Navigation in Real Environments Through Habitat»* - ICPR 2020

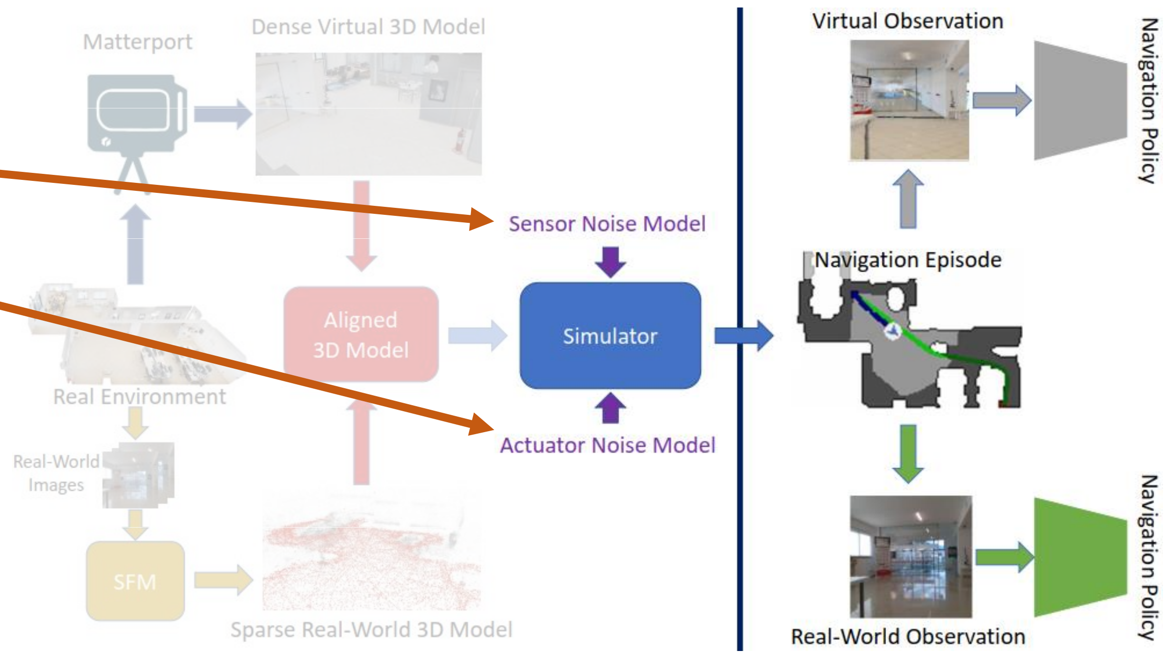
Introduce Real-World Inaccuracy

- Major difference between simulation and reality: presence of noise in perception and actions
  - Two noise models (Gaussian):
    - perception module (localization)
    - Action module (actuation)

- The simulator can produce navigation episodes with virtual and real-world observations
  - Running the simulator on the virtual environment and replacing the virtual observations with the real-world
  - 3DoF pose of the agent used to perform a fast retrieval of the nearest image in the real-world 3D model

  - camera coordinates (2D) + camera rotation, as unit vector

$$(u, v) = (cos\theta, sin\theta)$$



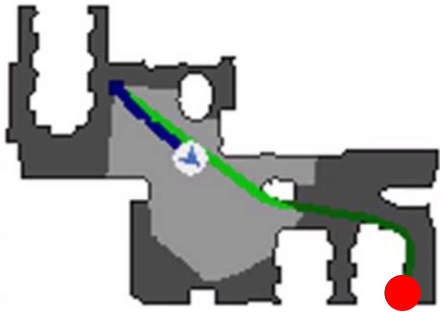
Rosano M., Furnari A., Gulino L., Farinella G.M. *«On Embodied Visual Navigation in Real Environments Through Habitat»* - ICPR 2020

Training Setup

- We formulated the navigation as PointGoal navigation task
  - (r, θ) polar coordinates, relative to the agent's position
  - Updated after each step

- At each timestep the agent receives the 256x256 RGB image + the goal coordinates

- Discrete actions:
  - *move straight by 0.25m;*
  - *turn left/right by 10°;*
  - *STOP*



- Trained using the PPO (Proximal Policy Optimization)[1] Reinforcement Learning algorithm
  - Success reward 2.5;
  - slack reward -0.01;
  - reward at each step $-(\, distance\ to\ goal_t - \ distance\ to\ goal_{t-1}) + slack$

- We used the RGB *DD-PPO* RL model [2], pretrained on the *Gibson*[3] and *Matterport*[4] datasets
  - *SE-ResNeXt50* [5] visual encoder + 2 layers 512-dim *LSTM* [6]
  - The visual embedding is concatenated to the information about the goal coordinates and the previous action performed, then fed to the LSTM

[1] Schulman et al., Proximal Policy Optimization Algorithms». In CoRR 2017
[2] Wijmans et al., «DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames». In ICLR 2020
[3] Xia et al., « Gibson Env: real-world perception for embodied agents». In CVPR 2018
[4] Chang et al., «Matterport3D: Learning from RGB-D Data in Indoor Environments». In 3DV 2017
[5] Hu et al., «Squeeze-and-Excitation Networks». In CVPR 2018
[6] Hochreiter et al., «Long short-term memory». In Neural computation 1997

Evaluation

- We randomly sampled 1000 navigation episode to evaluate the navigation episodes
  - Too easy episodes were discarded

- We split the real-world images into training and test sets of equal size, with images uniformely distributed in both sets

- An episode is considered successful if the agent calls the STOP action within 0.20m from the goal, or unsuccessful otherwise

- Navigation performance metrics:
  - SPL;
  - Success Rate (SR)

$$\text{SPL} = \frac{1}{N} \sum_{i=1}^{N} S_i \frac{l_i}{\max(p_i, l_i)}$$

where, l_i = length of shortest path between goal and target for an episode

p_i = length of path taken by agent in an episode

S_i = binary indicator of success in episode i

Rosano M., Furnari A., Gulino L., Farinella G.M. *«On Embodied Visual Navigation in Real Environments Through Habitat»* - ICPR 2020

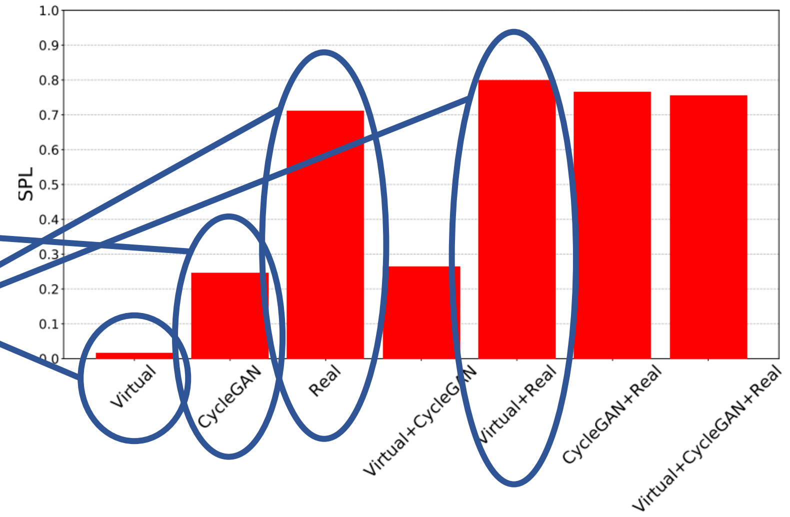
Results - Virtual to Real Policy Transfer

- We trained the models on different combination of virtual and real observations, and tested on real-world observations
- A+B indicates that the model was first trained on A, then fine-tuned on B
- CycleGAN as domain adaptation baseline
    - Trained for 50 epochs on 5k virtual images and 5k real-world images

VIRTUAL TO REAL POLICY TRANSFER

Training Stages	SPL	Success rate	Avg. dist. from goal (meters)	# training frames (Million)
Virtual	0.0160	0.022	7.9722	2.4
CycleGAN	0.2464	0.3310	4.6065	2.4
Virtual+CycleGAN	0.2648	0.3410	4.7535	1.2+1.2
Real	0.7112	0.8590	0.7709	2.4
Virtual+real	0.8001	0.9700	0.2493	1.2+1.2
CycleGAN+real	0.7665	0.8880	0.5219	1.2+1.2
Virtual+CycleGAN+real	0.7553	0.9360	0.3313	1.2+1.2+1.2

- The model trained only with virtual images achieved very limited performance
- The unsupervised adaptation with CycleGAN has led to significative better results
- Training the agent with real observations allows to obtain major performance improvements and pre-training the model with virtual observations allows to obtain additional improvements



Rosano M., Furnari A., Gulino L., Farinella G.M. *«On Embodied Visual Navigation in Real Environments Through Habitat»* - ICPR 2020

Results – Actuator and Sensor Noise

- We investigate the influence of actuator and sensor noise on visual navigation
- Do navigation models become more robust if trained in the presence of sensor/actuator noise?

- Two approaches:
    - Train the model without noise, test it with noise
    - Train and test the model with the same amount of noise

- The model trained and tested without noise obtained a good performance. However, even adding small amounts of noise during test degrades performance.

- Interestingly, part of the gap in performance is generally recovered by training the models in noisy settings

- Models trained without noise tend to terminate unsuccessful episodes at a greater distance than models trained with noise
    - Most of the episodes fail in the final part, near the goal

SENSOR AND ACTUATOR NOISE WITH VIRTUAL OBSERVATIONS

Sensors noise	Actuators noise	Trained with noise	SPL	Success rate	Avg. dist. from goal (meters)
No	No	No	0.9127	0.9910	0.1291
Small	No	No	0.8173	0.8910	0.1581
		Yes	0.8658	0.9380	0.1065
Medium	No	No	0.5075	0.5660	0.2404
		Yes	0.7114	0.7910	0.1554
Large	No	No	0.1552	0.1870	0.4909
		Yes	0.3643	0.4130	0.2577
No	Small	No	0.9092	0.9890	0.1171
		Yes	0.9073	0.9820	0.0956
No	Medium	No	0.8903	0.9700	0.1432
		Yes	0.8805	0.9740	0.1150
No	Large	No	0.8043	0.8860	0.2337
		Yes	0.8381	0.9340	0.2322
Small	Small	No	0.8020	0.8740	0.1876
		Yes	0.8328	0.8950	0.1607
Medium	Medium	No	0.4537	0.5100	0.2675
		Yes	0.4715	0.5290	0.2712
Large	Large	No	0.1288	0.1620	0.5442
		Yes	0.2450	0.2790	0.4040

	Noise level		
	Small	Medium	Large
Localization noise	$0.20m$; 7°	$0.40m$; 15°	$0.80m$; 30°
Actuation noise	$0.05m$; 5°	$0.10m$; 10°	$0.20m$; 20°

Rosano M., Furnari A., Gulino L., Farinella G.M. *«On Embodied Visual Navigation in Real Environments Through Habitat»* - ICPR 2020

Conclusion

- We investigated the problem of transferring visual navigation policies trained in simulation to the real world
- We proposed a tool based on Habitat to train and evaluate entirely in simulation visual navigation policies on real observations and with realistic sensor and actuator noise
- Adaptation methods are much needed to obtain visual navigation policies able to generalize to the real world
- The proposed framework is a promising tool to assess and improve their generalization ability when deployed in real contexts

We publicly released the collected dataset, the code and additional videos

https://iplab.dmi.unict.it/EmbodiedVN

Thank you for your time!

On Embodied Visual Navigation in Real Environments Through Habitat

Marco Rosano[1,3], Antonino Furnari[1], Luigi Gulino[3], Giovanni Maria Farinella[1,2]

UNIVERSITÀ degli STUDI di CATANIA

orangeDEV BE INNOVATIVE

ICAR CNR