2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Robust Navigation with Cross-Modal Fusion and Knowledge Transfer

Wenzhe Cai†, Guangran Cheng†, Lingyue Kong, Lu Dong, Changyin Sun*

***Abstract*— Recently, learning-based approaches show promising results in navigation tasks. However, the poor generalization capability and the simulation-reality gap prevent a wide range of applications. We consider the problem of improving the generalization of mobile robots and achieving sim-to-real transfer for navigation skills. To that end, we propose a cross-modal fusion method and a knowledge transfer framework for better generalization. This is realized by a teacher-student distillation architecture. The teacher learns a discriminative representation and the near-perfect policy in an ideal environment. By imitating the behavior and representation of the teacher, the student is able to align the features from noisy multi-modal input and reduce the influence of variations on navigation policy. We evaluate our method in simulated and real-world environments. Experiments show that our method outperforms the baselines by a large margin and achieves robust navigation performance with varying working conditions.**

I. Introduction

While the SLAM-based traditional navigation approaches [1]–[3] enable robot with navigation skills, they largely depend on complicated manually-designed modules. It may become fragile in environments with dynamic objects, pose estimation errors, and low texture. The learning-based methods, with concise end-to-end training architecture and powerful deep neural networks, attract diverse types of studies in navigation problems. Many representative works are proposed in point-to-point navigation [4], [5], object-goal navigation [6], [7], visual-language navigation [8], [9]. Most works model the navigation problem as a Markov Decision Process (MDP) and refer to deep reinforcement learning (DRL) algorithms for decision-making. Considering the high sampling costs in the trial-and-error learning process and low data efficiency in DRL, it is expensive to train the robot in real world. However, the zero-shot transfer to reality can lead to unexpected or even dangerous consequences. Therefore, improving the generalization of learning-based navigation methods is important for bridging the sim-to-real gap.

In real-world navigation systems, robots are usually equipped with multiple sensors (lasers, cameras, IMUs). Accurate perception is a prerequisite to train a robust navigation policy. To deal with ubiquitous sensor noise, researchers point out that multi-modal observations provide complementary properties [10], [11]. Therefore, it is possible to detect and identify data misalignment and error with advanced multi-sensor data fusion methods. Another important class of methods skips the explicit denoising procedure but concentrates on improving the policy generalization. Lots of deep learning-based approaches [12], [13] use domain randomization to avoid overfitting the training set. This is proved to be an effective way to acquire more robust policies. However, domain randomization methods may corrupt the distributions of data in the original state space. Since the exploration ability is a bottleneck for deep reinforcement learning algorithms, it adds the difficulty for DRL methods to discover the most task-relavant patterns, which may lead to sub-optimal policies.

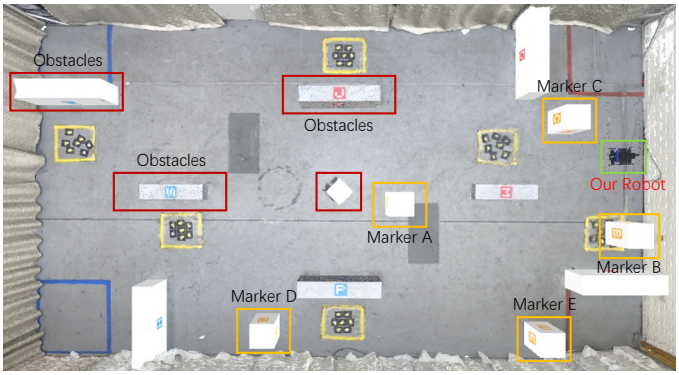

Fig. 1. A brief introduction of the navigation task. The robot (green rectangle) plans a navigation route to goals A,B,C,D,E in order. The initial position of of robot and goals (yellow rectangles) are randomized at each episode. Other obstacles are marked in red.

To deal with sensor noises and learn a generic navigation policy, we propose a cross-modal fusion network that is capable of extracting both task-relevant and noise-sensitive features. We train a robust navigation policy with expert knowledge transfer. This is completed by a teacher-student distillation framework. Concretely, the teacher module is responsible for generating an expert navigation policy. We train an RL agent in the ideal simulated environment (without sensor noise) as the expert policy without additional prior knowledge. The student module is designed to imitate the teacher's ways to perceive and react in the ideal environment to perform navigation with the existence of sensor noise. By introducing such a two-phase training procedure, the student module is able to better understand the dynamics and environment transitions, then learns a more generalized navigation policy. In addition, the student module also learns a transform function to restore the original signals from multi-modal inputs to resist the noise. To encourage learning

† Equal Contribution
* Corresponding Author

Wenzhe Cai, Guangran Cheng, Lingyue Kong and Changyin Sun are with the school of Automation, Southeast University, Nanjing, 210096, China, Lu Dong is with the school of Cyber science and Engineering, Southeast University, Nanjing, 210096, China, Emails: {wz_cai,chenggr,lingyuekong,ldong90,cysun}@seu.edu.cn.

This paper is supported by the National Key Research and Development Program of China under Grant 2018AAA0101400, the National Nature Science Foundation of China under Grant 61821004, 62173251, and the Nature Science Foundation of Jiangsu Province of China under Grant BK20202006.

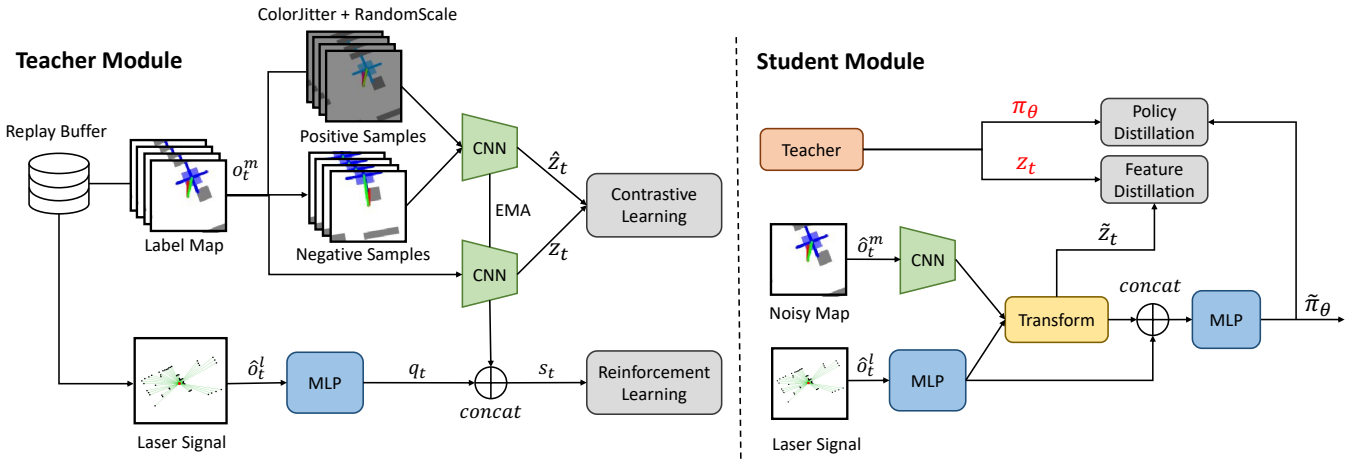

Fig. 2. Teacher-Student Distillation Framework of our method. To train a teacher policy, we use label map observation o_t^m and noisy laser observation $\hat{o}_t^l$ as policy inputs. In order to extract a useful representation that is sensitive to the agent pose information, we add a contrastive objective to widen the gap between two extracted features $\hat{z}_t$ and z_t of the augmented images, which are different in relative position information. Similar to MoCo [14], we use exponential moving average (EMA) to update one of the encoders. With the noisy input of map $\hat{o}_t^m$ and laser $\hat{o}_t^l$, the student is trained with supervised learning and tries to imitate the output from the encoder network z_t and policy network π_θ of the teacher. Finally, the student is able to learn a transform function to align the feature $\tilde{z}_t$ to z_t and imitate the teacher behavior with $\tilde{\pi}_\theta$.

of discriminative representations with respect to sensor noise, we introduce an auxiliary contrastive objective for teacher module. As pointed out in [15], self-localization accuracy is important for navigation. In this work, we focus on the pose estimation error and discuss the generalization of our method in navigation scenarios as shown in Figure 1. We evaluate our method both simulated and physical environments. The results show that our method improves the generalization with a large margin with varying working conditions (speed limit, noise scale, scene layout).

In a nutshell, our contributions are summarized as follows:

- We propose an expert knowledge transfer approach for robust robot navigation, which is completed with a teacher-student distillation framework.
- By introducing contrastive objective and feature distillation, our proposed cross-modal fusion approach can learn noise-sensitive representation and compensate for sensor noise.
- Empirical studies show the robustness of our navigation policy across various working conditions.

II. Related Work

A. DRL-based Navigation

Recently, many works employ DRL for navigation tasks. Some methods follow the traditional SLAM frameworks but replace crucial components (e.g. localization, map building, path-planning modules) with deep learning methods. This helps realize long-distance navigation in complex environments [16], [17]. Other methods collect data in simulators and directly learn monolithic navigation policies without planning guidance [18], [19]. To improve the data efficiency, task-relevant auxiliary objectives are introduced in addition to the origin RL objective [20]–[22]. Similar to the prior works, we also model navigation as an RL problem, but we use a two-phase training framework that contains both reinforcement learning and supervised learning. This outperforms the agent trained only with reinforcement learning.

B. Multi-Modal Fusion

In realistic navigation problems, an important issue is how to represent and fuse the multi-sensor observations in an appropriate way. Some methods stack the feature maps along the depth or concatenate them as flattened vectors before they are advanced to the outputs of the downstream task [23], [24]. To explore the relative informativeness of different sensing modalities, Mixture of Experts (MoE) approaches process each feature map by domain-specific networks and model their weights explicitly [25]. We propose a multi-modal fusion module to compensate for the noise. Different from the prior works, our fusion module explicitly learns a transform function and restored the original features.

C. Generalization of Deep Reinforcement Learning

Domain randomization is an important technique to improve the generalization of DRL methods. With randomized dynamical attributes in simulation, the agent can learn the policy in a diverse set of samples, which this helps learn a more robust policy [26]–[28]. Data augmentation can be regarded as a specific type of domain randomization method and many works refer to it for improving generalization [29], [30]. For example, by replacing the background texture, SODA [29] increases the success rate on pixel-to-control tasks. Policy distillation is also an effective way to improve generalization by extracting knowledge from an experienced expert [31], [32]. Our work is inspired by policy distillation methods. But we introduce a contrastive loss and this is proved to be necessary for better navigation policy.

III. Approach

A. Problem Formulation

We formulate the robot navigation problem as a Markov Decision Process (MDP) defined by a tuple (S, A, R, P, γ),

where S, A, P, R represents the state space, the action space, the reward function and, the transition function respectively. γ is a scalar and represents the discount factor. An optimal policy π^* aims to maximum the discounted culmulative reward $G = E_{s\sim\pi}[\sum_{t=0}^{T} \gamma^t R(s_t, a_t)]$, where a_t is sampled w.r.t the policy $\pi(s_t)$. Generally, instead of making decisions by the underlying state s_t, the agent can only access the sensor observations in navigation problems. Thus, it becomes a Partially Observable Markov Decision Process (POMDP). Concretely, we consider two types of observations for the navigation task, which are laser signal $o^l \in \mathbb{R}^n$ and egocentric occupancy map $o^m \in \mathbb{R}^{c\times h\times w}$. Here we focus on how to perform robust navigation skills, therefore, we assume the global occupancy map is already constructed with existing methods. At each time step t, the agent estimates its own pose $\hat{p}_t = (\hat{x}_t, \hat{y}_t, \hat{r}_t)$, where $\hat{x}_t$ and $\hat{y}_t$ represent the corresponding coordinates, and $\hat{r}_t$ represents the yaw angle. The egocentric occupancy map is a cropped global map with the center at $\hat{p}_t$. Note that the noise in pose estimation error makes the occupancy map o^m mistakenly reflect the surroundings, which tests the perceiving ability of the agent. The navigation targets are indicated by five goal coordinates denoted as $g_{A,B,C,D,E} = \{(x_A, y_A), ..., (x_E, y_E)\}$. By taking in the information from the noisy map, goal coordinates, and laser signals, the navigation policy needs to learn a mapping from observations to the robot control command $a_t \in \mathbb{R}^d$.

B. Cross-Modal Fusion with Distillation

Instead of training a navigation policy end-to-end with reinforcement learning, we propose a teacher-student distillation architecture for cross-modal fusion and robust policy learning. This is achieved by an teacher reinforcement learning agent and a supervised student agent. The illustration of our approach can be referred in Fig. 2.

Teacher Module: The teacher module learns navigation skills in environments without noise. Under such circumstances, the teacher is able to learn a near-perfect navigation policy π_θ under the deep reinforcement learning framework. To train that policy, we use a label egocentric map o_t^m without pose estimation noise, together with the laser signal $\hat{o}_t^l$ as the input observation space. The egocentric map is processed by a 3 layers of convolutional neural networks (CNN) and the laser signal is processed by 2 layers of fully-connected network (FC). Denote the feature from egocentric map as z_t and the feature from laser as q_t. The concatenated vector (z_t, q_t) are regarded as the state s_t for the actor-critic networks. We train the teacher module with *PPO* [33]. To faciliate efficient training, we design a dense reward signal defined as follows:

$$r_t = r_t^{step} + r_t^{col} + r_t^{goal} + \Delta d_t^p \tag{1}$$

where $r_t^{step} = -0.01$ is a constant penalty, $r^{col} = -0.05$ is collision penalty, $r^{goal} = 4.0$ when robot reaches a marker, Δd_t^p is the change of Eculidean distance to goal.

The DRL agent itself guarantees no improvement of generalization, especially for an agent trained with ideal conditions. Therefore, we introduce a contrastive objective to learn a discriminative feature representation that is sensitive to pose estimation error. We construct positive and negative pairs for egocentric maps and use *InfoNCE* [34] loss for optimization. Two augmented images from the same egocentric map o_t^m are regarded as positive pairs (q, k_+) while the rest in the mini-batch compose the negative pairs (q, k_-). In order to train a representation that can capture the pose information, we use *ColorJitter* and *RandomScale* as the data augmentation methods. Thus the network will learn to neglect the variations in size and color but focus on the relative pose information on the map. Similar to *MoCo* [14], we use a momentum encoder to avoid mode collapse. The auxiliary objective is proved to be crucial and a detailed ablation study are described in the experiment part. The *InfoNCE* loss is defined as:

$$\mathscr{L}^{InfoNCE} = -\mathbb{E}[log \frac{exp(q\cdot k_+/\tau)}{\sum_{j=0}^{B} exp(q\cdot k_-/\tau)}] \tag{2}$$

τ is a temperature hyper-parameter and we set $\tau = 0.25$ here. We balance the importance of the auxiliary objective with a coefficient β and the final loss function of the teacher module is defined as follows:

$$\mathscr{L}^{teacher} = \mathscr{L}^{PPO} + \beta^t \cdot \mathscr{L}^{InfoNCE} \tag{3}$$

We set $\beta^t = 0.2$ in our experiments. And $\mathscr{L}^{PPO}$ is the prposed reinforcement learning objective in [33]. we run 4 parallel environments with 2M steps to train the teacher module.

Student Module: The student module contains a multi-modal fusion network and a policy network which are both trained with supervised learning. We expect the student to imitate the teacher policy as similarly as possible. To that end, we introduce a feature distillation objective as well as a policy distillation objective. Specifically, the multi-modal fusion network takes in the noisy egocentric map $\hat{o}_t^m$ and the laser signal $\hat{o}_t^l$. Information from the map $\hat{o}_t^m$ is processed with 3-layers of CNN and the laser signal is processed with 2-layers of FC. Denote the feature of two modals as $\hat{z}^t$ and $\hat{q}^t$. Since the teacher module has been able to extract task-relevant features, we train the student network to learn a multi-modal fusion function to predict the teacher feature z_t. The fusion is completed by a transform module which contains 2 layers of FC. For simplicity, we use FC layers to fuse the information and align with the teacher feature, but any advanced architecture can be equipped here. As we train the agents in simulators, it is feasible to collect the paired observation data, one with noise and the other not. Therefore, at each time step, we can calculate the discrepancy between the teacher's representation and the student representation. Denote the transform layer as T_θ, the feature distillation objective is defined as negative cosine similarity:

$$\mathscr{L}^{FD} = -\mathbb{E}[\frac{z_t \cdot T_\theta(\hat{z}^t, \hat{q}^t)}{||z_t|| \cdot ||T_\theta(\hat{z}^t, \hat{q}^t)||}] \tag{4}$$

Although we train a feature transformation function, tiny difference between the teacher representation and the student representation can be magnified by policy network. As a

TABLE I

HYPER-PARAMETERS DETAILS OF OUR METHOD

	Teacher Module	Student Module
Optimization	Reinforcement Learning	Supervised Learning
Algorithm	PPO	-
Parallels	4	2
Training Steps	2M	1M
Optimizer	Adam	Adam
Discount Factor	0.99	-
GAE-λ	0.95	-
PPO-CLIP	0.15	-
nsteps	256	256
nepochs	2	2
nminibatch	4	2
learning_rate	4e-4	2e-4
lr_rate_decay	Linear	Linear

result, we also use a policy distillation objective which is defined as follows:

$$\mathscr{L}^{PD} = \mathbb{E}[KL(\pi^t_\theta(s_t)||\pi^s_\theta(s_t))] \tag{5}$$

We optimize a weighted loss function which is defined as:

$$\mathscr{L}^{student} = \alpha^s \cdot \mathscr{F}^{FD} + \beta^s \cdot \mathscr{F}^{PD} \tag{6}$$

We use $\alpha^s = 0.25$ and $\beta^s = 0.75$ in our experiments. A detailed table of hyper-parameters are listed in Table I.

IV. EXPERIMENT

A. Experiment Setup

The training environment is at the size of $8.08m \times 4.48m$. The laser signal o^l_t covers the detection range from -135 degree to 135 degree with $N = 60$ rays. The egocentric map o^m_t reflects the robot surroundings in occupancy map at the pose $\hat{p}_t$ in $2.56m \times 2.56m$. The control command $a_t = (v^x, v^y, \omega)$ are 3-dim continuous vectors including the linear speed and angular speed along the robot Cartesian coordinate system. To evaluate the generalization of our proposed method, we consider different variations between training and testing, including speed limit, noise scale and scene layouts. Specifically, during the training, the maximum of speed is $(2m/s, 2m/s, \frac{\pi}{4} rad/s)$. Without loss of generality, we consider an episodic shift $\delta^{shift} \sim U(-0.5, 0.5)$ and uniform noise $\delta^p_t \sim U(-0.1, 0.1)$ as pose estimation error. The laser signal is also injected with uniform noise $\delta^l_t \sim U(-0.1, 0.1)$. δ^{shift} stays fixed during one episode but different across episodes. The uniform noise δ^p,δ^l is variant at each time step. Five navigation goals are represented by (x, y) coordinates and the position of goals are randomized at each episode. During the training, the robot is requested to navigate to goals following the alphabet order in 500 steps, roughly equal to 20 seconds of clock time. The testing time is set to 60 seconds. It is worth to mention that the control mode between training and testing is different: In training, the interaction between robot and environment obeys MDP setting, but in testing, all the interaction are happened real-time, not only the control command but the control frequency also influences the performance. This result in slightly different state transition functions between training and testing. An overview of environment structures in training and testing are shown in Fig 3.

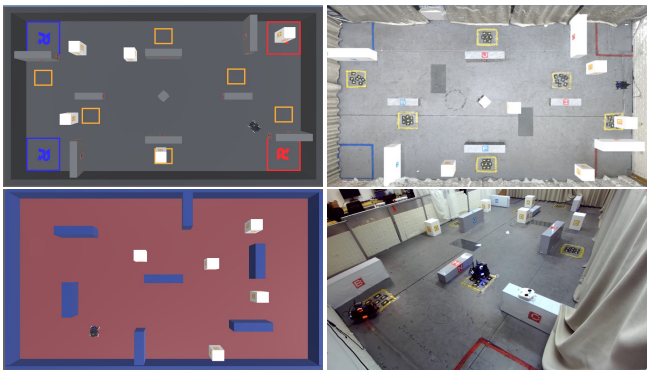

Fig. 3. Training and Testing environments in our experiments. All the training is completed in the top-left environments. We test the generalization across speed limit and noise in the second top-right environment. The left-bottom environment is used for testing the generalization across scenes. The last is an overview of our robot navigation environment in the real world.

B. Evaluation Metrics

We consider the following metrics to evaluate the generalization of different navigation methods:

- **Success Rate:** The average count of episodes where the robot successfully navigates to the 5 goals in order within 60 seconds.
- **Activation Count:** The average count of the activated goals.
- **Navigation Time:** The average clock time used for the navigation task.
- **Collision:** The average clock time of collision in an episode.

C. Baselines Methods

We implement three different approaches as the baselines, denoted as *Pose+Laser*, *Map+Laser*, *Teacher*. The details of the baselines are described as follows:

- **Pose+Laser**: An RL agent that concatenates the features extracted from pose estimation and noisy laser signal as states for policy inputs.
- **Map+Laser**: Different from the first method, it integrates the surrounding information and the robot pose estimation into an egocentric map. And it uses the extracted feature from CNN to replace the pose feature for policy inputs.
- **Teacher**: Different from the former two methods, it learns the navigation policy in an ideal environment without sensor noisy. We directly test the zero-shot performance in the noisy environment.

D. Generalization Across Speed Limit

In this experiment, we consider two speed limit settings $(1m/s, 1m/s, \frac{\pi}{4} rad/s)$, $(0.5m/s, 0.5m/s, \frac{\pi}{6} rad/s)$. We use the same noise settings in training. Changing the speed limit raises the problem of dynamics mismatch and results in a new MDP which owns a different transition function. Although two MDPs share common task-relevant attributes, the robot will encounter novel states that have not appeared. We list the performance with two settings in Table II. Only the *Map + Laser* achieves the compatible performance in

TABLE II

PERFORMANCE ON SPEED GENERALIZATION.

$v_x = 1m/s, v_y = 1m/s, \omega = \pi/4\ rad/s$				
Methods	Success(%)	Activation	Collision(s)	NavTime(s)
Pose + Laser	61.8	3.54	10.14	37.65
Map + Laser	90.6	4.72	8.37	26.48
Teacher	74.1	4.34	5.73	35.06
Ours	**98.1**	**4.96**	**2.01**	**21.73**
$v_x = 0.5m/s, v_y = 0.5m/s, \omega = \pi/6\ rad/s$				
Methods	Success(%)	Activation	Collision(s)	NavTime(s)
Pose + Laser	28.2	2.46	8.45	54.81
Map + Laser	84.1	4.46	10.51	45.18
Teacher	42.6	3.31	4.27	50.43
Ours	**88.1**	**4.53**	**2.34**	**43.83**

TABLE III

PERFORMANCE ON NOISE GENERALIZATION.

$\delta^{shift} \sim U(-0.25, 0.25), \delta_t^p \sim U(-0.05, 0.05), \delta_t^l \sim U(-0.05, 0.05)$				
Methods	Success(%)	Activation	Collision(s)	NavTime(s)
Pose + Laser	71.6	4.19	10.96	35.01
Map + Laser	95.7	4.82	3.75	**21.75**
Teacher	94.3	4.84	4.48	24.91
Ours	**97.3**	**4.90**	**2.44**	22.31
$\delta^{shift} \sim U(-0.75, 0.75), \delta_t^p \sim U(-0.15, 0.15), \delta_t^l \sim U(-0.15, 0.15)$				
Methods	Success(%)	Activation	Collision(s)	NavTime(s)
Pose + Laser	34.2	2.84	13.34	48.57
Map + Laser	64.7	3.86	7.26	37.61
Teacher	34.1	2.64	4.16	49.87
Ours	**81.2**	**4.59**	**3.45**	**32.48**

success rate and activation count, but it fails in collision avoidance, while our method outperforms all the baselines with a large margin in all the metrics.

Note that the cross-modal fusion module is designed to compensate for the sensor noise, but also shows better generalization ability with respect to the speed limit. It implies that policy distillation is beneficial to learn a better representation for environment dynamics.

E. Generalization Across Noise

In this experiment, we fix the speed setting as $(1m/s, 1m/s, \frac{\pi}{4} rad/s)$ and discuss the influence of the noise with different scales. The results are reported in Table III. Two different noise settings are considered here: In the easy scenarios, we set the episode shift $\delta^{shift} \sim U(-0.25, 0.25)$ and uniform noise for pose estimation error as $\delta_t^p \sim U(-0.05, 0.05)$. The uniform noise for laser signal is also set to $\delta_t^p \sim U(-0.05, 0.05)$. Most approaches achieve good performance (90+% success rate) under this settings. But when we increase the noise, i.e. $\delta^{shift} \sim U(-0.75, 0.75), \delta_t^p \sim U(-0.15, 0.15), \delta_t^l \sim U(-0.15, 0.15)$, the performance of baseline methods degrade dramatically, especially for the teacher module (nearly 60% drop on success rate). Without training in noisy environments, the teacher policy cannot handle situations with unseen states. Our method still maintains a satisfying performance, outperforming the best baselines *Map+Laser* with 16.5% on success rate and nearly a half collision time. By imitating the teacher representation and policy, we reckon that it learns the transform function to restore the original feature by incorporating the information from two modals. To verify our hypothesis, we visualize the features before and after the transform layers in the student module and compare it with the label feature extracted from the teacher module. T-sne visualization of 3 types of features are shown in Fig 4. Before feeding into the transform layer, the noisy feature (orange dots) is far from the label features (blue dots). And the aligned feature (green dots) is close to the label features. This shows that our cross-modal fusion approach can transform the noisy feature into the labels' neighborhood and reduce the influence of sensor noise.

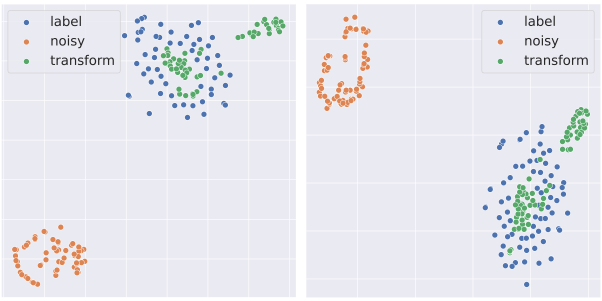

Fig. 4. T-sne visualization of the embedding feature from the Label Map, the Noisy Map, and the Transformed Feature with Multi-Modal Fusion, which are represented by blue,orange,and green dots respectively. Our proposed multi-modal fusion method greatly reduces the gap to the label map.

F. Generalization Across Scene

In real-world applications, robots are often requested to work in new environments. Therefore, zero-shot generalization across different scenes is an important ability for a navigation policy. We test the performance of all the methods in an unseen environment with a different layout (as shown in Fig 3). The speed limit is set to $(v_x = 1m/s, v_y = 1m/s, \omega = \pi/4\ rad/s)$ and the noise scale is set to $\delta^{shift} \sim U(-0.5, 0.5)$, $\delta_t^p \sim U(-0.1, 0.1)$, and $\delta_t^p \sim U(-0.1, 0.1)$. We report the metrics in the test scene in Table IV. Since the *Pose+Laser* is simply overfitting the training environment, it fails in the testing environment and gets zero success rate. The spatial information is crucial for navigation skills. Even though the agents are trained with only one scene, the other three methods all show generalization ability to novel scenes. This encourages us that the egocentric map is an appropriate state space design for navigation tasks. And our method shows 80% success rate on the new scene. We believe that the proposed method can achieve better generalization performance by training with more scenes and this is a direction for our future work. By then, it can serves as a motion planning module and directly embed into the SLAM framework.

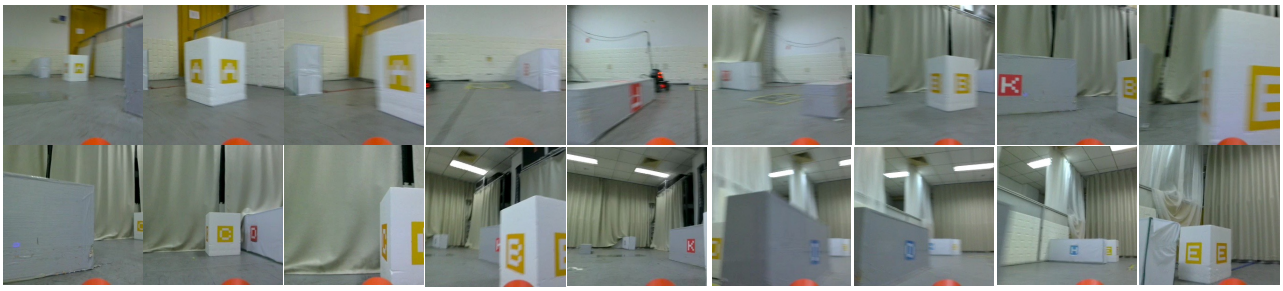

Fig. 5. First-person view of navigation trajectory. We illustrate key frames and the robot successfully reaches A,B,C,D,E in order.

TABLE IV

PERFORMANCE ON NOVEL SCENE.

Performance on Test Scene				
Methods	Success(%)	Activation	Collision(s)	NavTime(s)
Pose + Laser	0.0	-	-	-
Map + Laser	72.2	4.19	10.02	**31.98**
Teacher	61.5	3.68	10.48	45.28
Ours	**80.2**	**4.46**	**6.82**	33.67

G. Sim-to-Real Generalization

The real-world experiments are implemented on a DJI RoboMaster EP robot with an RPLiDAR S2 and a nvidia Xavier NX module. An overview of the EP robot is shown in Fig 6. in the real world experiment, the layout is the same as the training environment. To get the pose information,we use the point-cloud matching with an existing map to acquire the robot pose. Mismatch can happen, so it will lead to a different distribution of pose estimation noise. A first-person view of the navigation trajectories are shown in Figure 5. The robot successfully reaches A,B,C,D,E. For more information, please refer to the attached video.

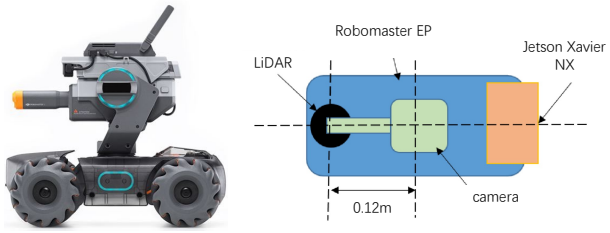

Fig. 6. Overview of the Robomaster EP robot.

H. Ablation Study

We make three ablation studies to understand the contribution of each component. In the first ablation, we remove the transform layers but directly use the CNN features of the egocentric map to finish the feature distillation. We show the negative cosine similarity during the training in Fig 7. And we notice that without the laser signal, the similarity converges at around 0.4, but our multi-modal fusion can get more than 0.6, this proves that laser signal can provide additional information for feature restore and alignment. In the second ablation, we remove the contrastive objective in teacher module. In the third ablation, we remove the feature distillation objective in the student module. We test

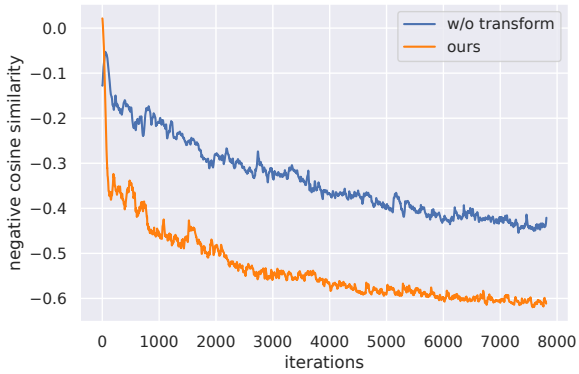

Fig. 7. Negative Cosine Similarity in training.

TABLE V

ABLATION STUDY WITHOUT FEATURE DISTILLATION OR WITHOUT CONTRASTIVE OBJECTIVE.

$\delta^{shift} \sim U(-0.75,0.75), \delta_t^p \sim U(-0.15,0.15), \delta_t^l \sim U(-0.15,0.15)$				
Methods	Success(%)	Activation	Collision(s)	NavTime(s)
w/o ft. distill	72.5	4.15	**3.32**	36.95
w/o cont.	56.1	3.88	4.01	42.71
Ours	**81.2**	**4.59**	3.45	**32.48**

the navigation performance of the above methods in the settings same as section E. The results are shown in Table V. Notably, without contrastive objective, the teacher still perfectly solves the navigation task in the ideal environment, but it fails to teach a good student. And the performance also decreases w/o feature distillation. All the ablation studies suggest that selecting an proper representation space is an essential problem.

V. CONCLUSION

We propose a multi-modal fusion method to incorporate the information from laser signal and map information. By training in a teacher-student distillation framework, our proposed CMFD method is able to compensate for the pose estimation noise and greatly improve the generalization capability across a diverse of working conditions. With sim-to-real experiments, we prove that our method is suitable for realistic navigation problems. But in our work, we haven't consider the dynamic obstacles. How to represent the locomotion information about the obstacles and incorporate a real-time mapping module for obstacle avoidance and navigation is one of our concerns for future works.

REFERENCES

[1] S. M. LaValle, "Planning algorithms," 2006.

[2] S. Thrun, "Probabilistic robotics," *Commun. ACM*, vol. 45, pp. 52–57, 2002.

[3] B. P. Wrobel, "Multiple view geometry in computer vision," *Künstliche Intell.*, vol. 15, p. 41, 2001.

[4] E. Wijmans, A. Kadian, A. S. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," in *ICLR*, 2020.

[5] X. Zhao, H. Agrawal, D. Batra, and A. G. Schwing, "The surprising effectiveness of visual odometry techniques for embodied pointgoal navigation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16 107–16 116, 2021.

[6] Y. Qiu, A. Pal, and H. I. Christensen, "Learning hierarchical relationships for object-goal navigation," *ArXiv*, vol. abs/2003.06749, 2020.

[7] D. S. Chaplot, D. Gandhi, A. K. Gupta, and R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *ArXiv*, vol. abs/2007.00643, 2020.

[8] H. Wang, W. Wang, W. Liang, C. Xiong, and J. Shen, "Structured scene memory for vision-language navigation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8451–8460, 2021.

[9] S. Chen, P.-L. Guhur, C. Schmid, and I. Laptev, "History aware multimodal transformer for vision-and-language navigation," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=SQxuiYf2TT

[10] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[11] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *ECCV*, 2018.

[12] K. Arndt, M. Hazara, A. Ghadirzadeh, and V. Kyrki, "Meta reinforcement learning for sim-to-real domain adaptation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2725–2731.

[13] J. Matas, S. James, and A. J. Davison, "Sim-to-real reinforcement learning for deformable object manipulation," in *Conference on Robot Learning*. PMLR, 2018, pp. 734–743.

[14] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020.

[15] S. Datta, O. Maksymets, J. Hoffman, S. Lee, D. Batra, and D. Parikh, "Integrating egocentric localization for more realistic point-goal navigation agents," in *CoRL*, 2020.

[16] A. Faust, K. Oslund, O. Ramirez, A. Francis, L. Tapia, M. Fiser, and J. Davidson, "Prm-rl: Long-range robotic navigation tasks by combining reinforcement learning and sampling-based planning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5113–5120.

[17] A. Francis, A. Faust, H.-T. L. Chiang, J. Hsu, J. C. Kew, M. Fiser, and T.-W. E. Lee, "Long-range indoor navigation with prm-rl," *IEEE Transactions on Robotics*, vol. 36, no. 4, pp. 1115–1134, 2020.

[18] H.-T. L. Chiang, A. Faust, M. Fiser, and A. Francis, "Learning navigation behaviors end-to-end with autorl," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 2007–2014, 2019.

[19] H. Shi, L. Shi, M. Xu, and K.-S. Hwang, "End-to-end navigation strategy with deep reinforcement learning for mobile robots," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2393–2402, 2019.

[20] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu *et al.*, "Learning to navigate in complex environments," *arXiv preprint arXiv:1611.03673*, 2016.

[21] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu, "Reinforcement learning with unsupervised auxiliary tasks," *ArXiv*, vol. abs/1611.05397, 2017.

[22] J. Kulhánek, E. Derner, T. De Bruin, and R. Babuška, "Vision-based navigation using deep reinforcement learning," in *2019 European Conference on Mobile Robots (ECMR)*. IEEE, 2019, pp. 1–8.

[23] S. Chadwick, W. Maddern, and P. Newman, "Distant vehicle detection using radar and vision," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8311–8317.

[24] J. Dou, J. Xue, and J. Fang, "Seg-voxelnet for 3d vehicle detection from rgb and lidar data," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4362–4368.

[25] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1239–1285, 2020.

[26] P. Anderson, A. Shrivastava, J. Truong, A. Majumdar, D. Parikh, D. Batra, and S. Lee, "Sim-to-real transfer for vision-and-language navigation," in *Conference on Robot Learning*. PMLR, 2021, pp. 671–681.

[27] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[28] J. Choi, K. Park, M. Kim, and S. Seok, "Deep reinforcement learning of navigation in a complex and crowded environment with a limited field of view," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5993–6000.

[29] N. Hansen and X. Wang, "Generalization in reinforcement learning by soft data augmentation," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13 611–13 617, 2021.

[30] I. Kostrikov, D. Yarats, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," *ArXiv*, vol. abs/2004.13649, 2021.

[31] R. Traoré, H. Caselles-Dupré, T. Lesort, T. Sun, N. Díaz-Rodríguez, and D. Filliat, "Continual reinforcement learning deployed in real-life using policy distillation and sim2real transfer," in *ICML Workshop on "Multi-Task and Lifelong Reinforcement Learning"*, 2019.

[32] B. Zhou, N. Kalra, and P. Krähenbühl, "Domain adaptation through task distillation," in *European Conference on Computer Vision*. Springer, 2020, pp. 664–680.

[33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *ArXiv*, vol. abs/1707.06347, 2017.

[34] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *ArXiv*, vol. abs/1807.03748, 2018.