OCEANS 2020, Singapore, August 11-14, 2020

Eye Gaze Map as an Efficient State Encoder for Underwater Task Automation

Wenjun Xu
Robotics Research Center
Peng Cheng Laboratory
Shenzhen, China
xuwj@pcl.ac.cn

Jianfeng Wei
Robotics Research Center
Peng Cheng Laboratory
Shenzhen, China
weijf@pcl.ac.cn

Renyou Yang
Robotics Research Center
Peng Cheng Laboratory
Shenzhen, China
yangry@pcl.ac.cn

Aidong Zhang
Robotics Research Center
Peng Cheng Laboratory
Shenzhen, China
zhangad@pcl.ac.cn

***Abstract*—ROVs are the primary tool used in marine environments for conducting dexterous manipulation tasks and the teleoperation of which requires experienced operators. In an attempt to ease the fatigue of the operators and to potentially improve task efficiency, we propose an image based framework to automate underwater routine tasks via imitation learning. In particular, we utilize the gaze information of the operator for extracting task-relevant information from the raw image input by an encoding network. We further demonstrate that the incoporation of the eye gaze encoder facilitates the training of the task policy which includes both maneuver and decision making.**

***Index Terms*—Gaze map, Imitation learning, ROV, Task automation, State encoder**

I. Introduction

Autonomous capture of sea animals such as sea cucumber is highly demanded, since manual capture performed by experienced divers presents high risk and is economically inefficient. However, control and task automation of underwater robots are challenging problems, primarily because of the lack of accurate robot dynamic models and the highly dynamic and unpredictable marine environments the robots are operated in. In addition, precise and real-time underwater localization are still open questions, which adds difficulty to the automation problems. Control and task automation frameworks for ROVs usually treats localization and control as two separate tasks: multi-modality sensors or object recognition networks are usually required [1] for object localization; as for decision and control, traditional model-based control [2], learning based control [3] and their hybrid [4] were investigated. In particular, vision based control or visual servoing predicts the target position from visual information. Current visual servo control for ROV, however, cannot cope with large external disturbances where the target is out of sight, Unless state estimation [5] and decision modules are incorporated.

We seek to solve the problem with inexpensive sensors like monocular camera and pressure sensor in an end-to-end fashion. End-to-end learning control, that outputs control and decision making commands directly from sensory input, receives wide research interest for applications in self-driving cars [6], [7] and robot manipulation [8], [9]. Recent effort in applying end-to-end sensory motor control methods for marine robotics can be seen in [10].

Traditional architectures of learning visualmotor control policies take raw images as the input, followed by convolutional neural networks for feature extraction. These architectures present low sample efficiency, weak intepretability and poor generalization to unseen environment and tasks. Gaze patterns have been studied by psychologists, neuroscientists and deep learning scientists [11], and shown to be important cues for decision making in a variety of tasks such as human-machine interactions [12] and self-driving cars [13]. Shi *et al.* improves the generalization capability of imitation networks by exploiting gaze map as an additional input and as a novel dropout mechanism. We intend to propose a different way of utilization of the gaze map to encode spatial location of the target in the image, which is highly relevant to the object targetting and capture task we are trying to solve. In contrast to [1], where the target locations are inferred by an object recognition network, extensive manual labeling is not required in our approach.

In this work, we consider the learning of a low dimensional state representation of raw images for underwater task automation, in particular, autonomous capture of sea cucumbers. The task requires the operator to first detect the object to catch, teleoperate the ROV to move towards the target object until they are close enough for the manipulator to catch. The latter two subtasks are considered here for automation: learn a maneuver policy for vision servo control; learn when to stop servoing and start closing the manipulator for capture (decision making).

We adopted the Learning from Demonstration (LfD) framework, which consists of (1) dataset collection; (2) training the eye gaze map for state representation; (3) training the policy network for maneuvering and decision making. Fig.1 shows the setup for dataset collection. An experienced ROV operator performed the task multiple times from vision feedback, where the initial state of the ROV is randomized and the location of the target is fixed. Images, depth of the ROV, eye gaze information of the operator and the joystick commands were recorded from the demonstrations. To simulate the terbulance environment, we apply external load to the ROV with a rigid rod manually, and 10% of the demonstrations were conducted with disturbance.

Key contributions of this work can be summarized as

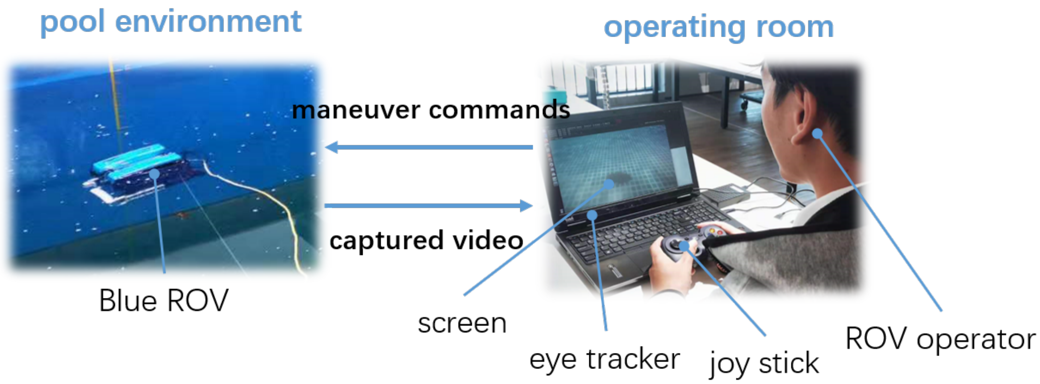

Fig. 1. Experimental setup of the data collection process. The Blue ROV equipped with monocular camera and depth sensor is deployed in the pool environment. A sea cucumber model is placed in the bottom of the pool as the object. An experienced ROV operator is asked to teleoperate the ROV to perform the object capture task with only vision feedback. A Tobii 4C eye tracker is mounted on the screen to record the gaze information.

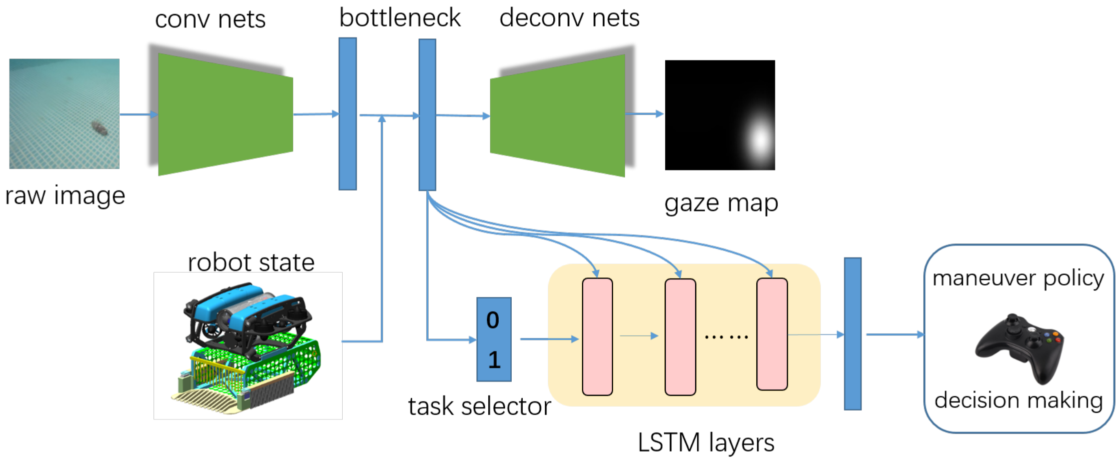

Fig. 2. Network architecture of the proposed method. The neural network consists of a gaze map prediction model that extracts task relavant information and a control/decision model that outputs either maneuver policy or decision signals acoording to the task selector. The gaze model is of a *conv- deconv* structure and the conv part is shared between the two tasks.

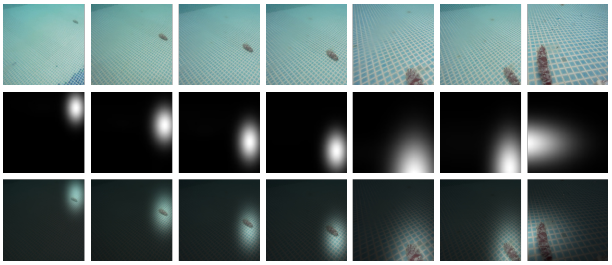

Fig. 3. The raw images (*First row*); the corresponding estimated gaze maps (*Second row*) and their overlaps(*Third row*)

follows:

- We train an eye gaze model via a convolution-deconvolution architecture to extract task-relavant features that encode spatial information of the targets. The trained network can predict and track the position of the object in both trained and unseen environments.
- We exploit the possibility of incorporating the trained gaze map into an end-to-end imitation learning framework, where the low-dimensional bottleneck layer of the gaze model is extracted as the input state to a LSTM network for both task segmentation and maneuver policy learning. The LSTM layer embeds sequential information and outputs robust policy even when the ROV is subject to external disturbances, which is particularly common in marine environments.
- The incorporation of the eye gaze information facilitates training of the imitation learning framework with reduced training time, but with comparative performance comparing to networks with raw image input. The proposed framework, without manual labeling process required, is also advantageous over traditional architectures that incoporate object recognition and visual servo control [1].

II. Methodology

Popular approaches to solve imitation learning problmes are Behavior Cloning (BC) and Inverse Reinforcement Learning (IRL) [14]. Here, we adopt the BC framework, for it's simple to apply. Behavior Cloning is formulated as supervised learning problems, where K completions of task demonstrations are collected, $\{\tau_1, \tau_2, ...\tau_K\}$, each consists of $N_k, k = 1, 2, ...K$ observations of state-action pairs $(s_i^k, a_i^k), i = 1, 2, ...N_k$. The dataset is then formed by $\mathcal{D} = \{(s_1, a_1), (s_2, a_2), ...\}$. The goal is to learn a policy $\pi_\theta(a_i|s_i)$, parameterized by θ, that can follow the policy from the demonastration as closely as possible, by minimizing the following loss,

$$\mathcal{L} = \mathbb{E}_{s_i, a_i \in \mathcal{D}} D_{KL}(\pi_\theta(a_i|s_i)||p(a_i|s_i)) \tag{1}$$

where we use KL-divergence to characterize the similarity between the learned policy and the expert policy.

One problem of Behavior Cloning is that it requires large amount of data samples to be able to learn an adequate policy, and we propose a state encoding method with additional gaze information that extract the most salient features relavant to the task to improve sample efficiency. Hence, we introduce a mapping $\tilde{s} = f(s)$ to represent the state encoder to map the original image input to the gaze-supervised low-dimensional feature.

A. Task Demonstration

The setup of the dataset collection process is shown in Fig.2. We performed the sea cucumber capture task with BlueROV (Blue Robotics Inc), a consumer level ROV equipped with a monocular camera to provide visual information of the environment, depth sensor and IMU to provide its positional and motion information. The ROV is connected by a Fathom tether to a topside PCfrom which, the ROV operator can visualize and monitor the state of the ROV in real time through a GUI and tele-operate the ROV through a joystick. A plastic basket was designed to store the sea cucumber and was installed below the ROV. A pair of 1 DoF manipulators were attached at the front of the basket to "sweep" the sea cucumber into the basket when the ROV approached the target. An eye tracker device, Tobii 4C (Tobii Gaming), was mounted at the screen of the PC to record the operator's eye gaze data in real time (runnig at 90Hz). A data collection routine was developed with ROS and Ubuntu 16.04 to package the image and various sensor data in ros bag format.

We asked a well-trained ROV operator to perform the demonstration, who was provided with visual information from the ROV camera only without the sight of the task environment. A sea cucumber model was placed in the pool in advance to serve as the target. At the start of each demonstration, the BlueROV was deployed in the pool at different locations with respect to the target. We kept the pitch of the camera fixed. We used a rod to interrupt the motion of the ROV at random directions to simulate the environmental disturbances. In each demonstration, the ROV is likely to go through a searching-approaching-landing process, and once the ROV landed close enough to the target such that the target is in the "sweeping" region of the manipulator, the demo is regarded as successful. In total, 24 successful demonstrations were recorded, and within which, 3 are with disturbances.

B. Dataset Preparation

We first extract images $\mathbf{I}$, depth data h, eye gaze data $\mathbf{e} = (u, v)$ and the 4 dimensional joystick commands $\mathbf{J} = (J_1, J_2, J_3, J_4)$ which corresponds to longitudinal, lateral, yaw and ascend/descend motion respectively, from the recoreded ros bag. The data were aligned temporally with a reference frequency of 33Hz. We manually segmented each demo into three phases: searching, approaching and landing and kept the latter two for subsequent process. We prepared three datsets corresponding to 3 models: gaze model, manuever model and decision making model. Image data were resized into $3 \times 224 \times 224$ and normalized for each model.

Gaze model:As eye data is highly dynamic and sensitive, we applied several filtering procedures to stably extract the gaze fixation. As fixation is defined as the pause of gaze within a spatially limited region ($0.5 - 1$degree) for a minimum period of time ($80 - 120ms$) [15], we ran a mean average filter with a window size of 10 (approximately 100ms). We ran a statistics analysis for each window by calculating standard deviation sx, sy and Pearson's r r. Data samples with $sx > 0.02$ or $sy > 0.02$ or $r \geq 0.5$ or $r \leq -0.5$ were filted out. We fitted a 2D Gaussian distribution for each running window and resized to a 64×64 gaze map for the target of the gaze model. The input to the model is the raw image, which was cropped into two sub-images to potentially increase the percentage of the "small object" in the dataset. Small object detection is a challenging yet active research area in computer vision, and many methods exist to enhance the detection rate [16], [17]. We found the cropping technique to be effective for our dataset as shown in Fig. 3. We also apply other augmentation techniques by randomly shifting the brightness, constrast and gamma online for each image while training. We obtained a dataset of $30,000$ samples and we devided the dataset into training set and testing set with a ratio of $3:1$.

Manuever model: The manuever model takes the low-dimensional bottlenet feature of the gaze model, appended with the depth information as the input, where the depth is normalized into $[-1, 1]$. The target of this model is the 4-axis joystick commands, data in each axis were discretized into 5 clusters by Gaussian Mixture Modeling (GMM) as shown in Fig.4. Hence, the manuever model was formulated as a $5^4 = 625$ classification problem.

Decision making model: The decision model was formulated as a 2D classification problem: 0 as servoing and 1 as landing and the data were labeled manually. The input is the same as the manuever model and thus was processed similarly.

C. Neural Network Structures

The network architecture is illustrated in Fig.2. The gaze model takes raw images as the input and outputs the estimated gaze maps. Inspired by the VAE-GAN architecture [18], we employed the *conv-deconv* structure, where the skeleton

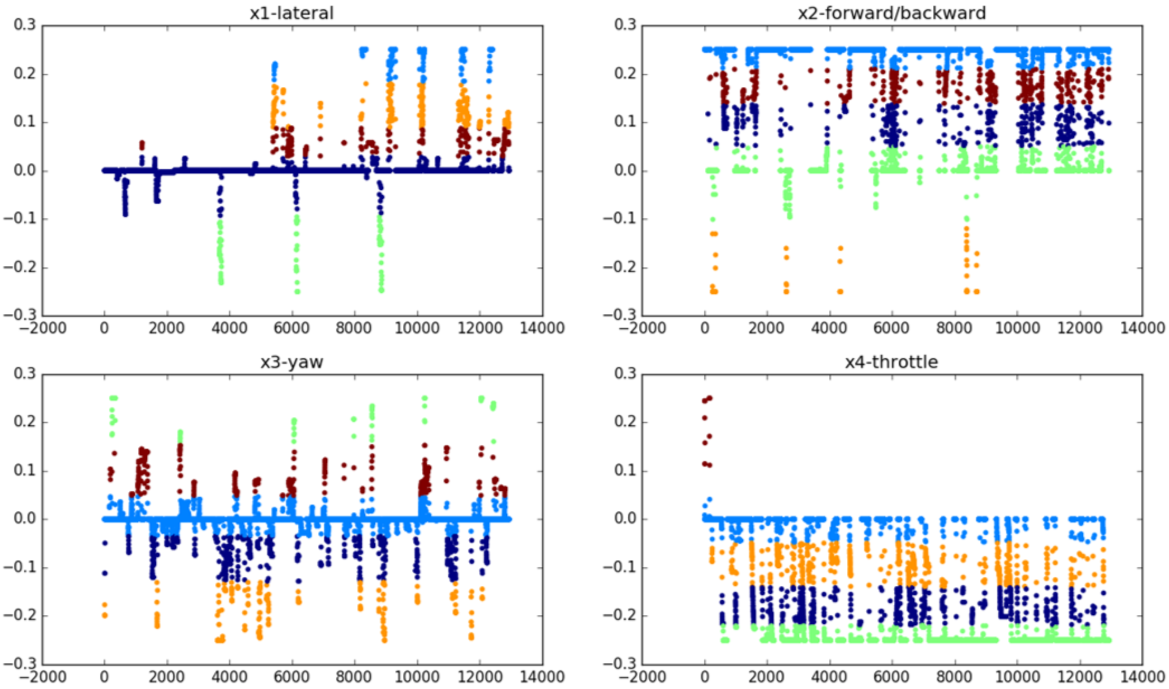

Fig. 4. Discretization of the joystick commands performed by GMM, where clusters were represented by different colors.

network is *Resnet18*. We were surprised to find that the gaze map stably encodes the position of the sea cucumber in the image, as demonstrated in Fig. 3. The bottleneck layer of the gaze model, together with the depth information of the ROV from the water surface, are the input to the control/decision making network, which is formed by LSTM layers. LSTM is expected to account for time history state information, to better assist decision making in robotic control tasks [19]. Particularly, localization techniques in dynamic underwater environment is not maturally developed and sensors like IMU that can tell time-history state information of ROV is subject to integration error, in such cases, however, visual information serves as an alternative to infer the history state of the ROV. Moreover, in target capturing task, lost tracking of the target is quite normal when the ROV is subject to external disturbances, traditional visual servoing controllers [10] without referring to the history states is invalid. The two subtasks, selected by a switch signal, share the same state representation input.

III. Experimental Results

Experimental results on the dataset collected from the pool environment can be seen in https://youtu.be/ed8WGyNUZTI. We detail the training and evaluation of the gaze model as well as the Imitation learning model in the following.

A. Evaluating the Gaze Model

The objective of the gaze model here is to investigate the possibility of encoding spatial information of the target in the image using gaze data, without extensive manual labeling process, that is indispensable for object detection networks [20]. We train our eye gaze model (model (a)) separately by minimizing the *KL divergence*, and compare the prediction performance of the proposed auto-encoder structure with two other models as illustrated in Fig. 5. Model (b) directly outputs the spatial location of the gaze point by minimizing the *mse* loss while training. Model (c) reconstructed the gaze map by an m component GMM with the same setting as in [21]:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{m} p(i) g_i(\boldsymbol{y}|\boldsymbol{x}) \quad (2)$$

$$L = -\ln\left\{\sum_{i=1}^{m} p(i) g_i(\boldsymbol{y}|\boldsymbol{x})\right\} \quad (3)$$

where $\boldsymbol{x}$ is the encoded feature, $\boldsymbol{y}$ is the target gaze point, $p(i)$ is the mixing coefficient and they sum up to 1: $\sum_{i=1}^{m} p(i) = 1$, $g(i)$ is a multivariate Gaussian with mean u_i and variance σ_i^2. We set $m = 4$ in the experiments after several rounds of trials. The loss is defined as the negative logarithm likelihood of 2 as in 3. The setting of the hyperparameters for training the 3 models are the same, where the learning rate is $lr = 0.001$, SGD optimizer are adopted with momentum of 0.9 and weight decay of 0.001.

Model (a) and (b) converges easily within less than 50 ephochs, while we have to early stop model (c) to obtain the results. Training, testing and prediction results are summarized in Table I. We quantify the similarity of the predicted gaze with the true target. For fair comparison, the prediction error for all 3 models is quantified by *mse*. For model (a) and (b), we extracted the point (x_{max}, y_{max}) that takes the maximum value of the reconstructed gaze map, which was represented by the red dot in the supplementary video, so that mse can be calculated. The proposed eye gaze model outputs the minimum mse as shown in Table I, so we adopt the same structure in the imitation model.

TABLE I
TRAINING AND PREDICTION PERFORMANCE OF THE 3 GAZE MODELS. TRAINING AND TESTING PERFORMANCE FOR THE 3 MODELS ARE QUANTIFIED BY THEIR LOSS FUNCTIONS INTRODUCED IN SECTION III-A. PREDICTION ERROR IS QUANTIFIED BY MEAN SQUARED ERROR FOR ALL 3 MODELS.

	Train	**Test**	**Prediction**
Model(a)	0.0146	0.5550	**0.0040**
Model(b)	0.0085	0.1466	0.0105
Model(c)	0.1676	1.3896	0.0809

B. Training the Imitation Network

We adopted a two layer LSTM with the latent space size as $512 + 1$ (512 from the encoded eye gaze information and 1 for the depth). We trained the network with a sequence of 8 time-steps and batch size of 32 for 50 epochs. RMSProp was adopted as the optimizer with initial learning rate of 0.005 and decay of 0.999. At test time, the entire sequence of a demonstration was fed into the network for prediction. Since the data samples in the two categories for the decision making datset are imbalanced, we applied random downsampling to the class with more samples for data balancing [22]. The prediction results are demonstrated in the supplementary video, where we overlay the original video with 2 dots to emulate the joystick commands. One nice property of the trained policy is that while the target is out of the screen due to disturbances, the policy can drive the ROV until the target is back to the screen, without substantial localization techniques and decision modules.

To illustrate the capability of the eye gaze map to encode low-dimensional task-relevant features, we compare the proposed architecture on the decision making task, with an state-of-the-art end-to-end network without eye gaze information, but instead uses a spatial softmax layer as an encoder [8]. The classification results are represented by confusion matrix and are shown in Fig. 6. We can observe that the performance of the two architectures are comparable, with the eye gaze model performing slightly better. Moreover, our model offers better interpretability.

IV. CONCLUSION AND FUTURE WORK

In this paper, we propose an end-to-end imitation network to automate routined ROV underwater manipulation tasks, where we incoporate eye gaze data to encode low-dimensional task relevant information from the original high dimension input. We successfully demonstrate that by an encoding network, the predicted eye gaze map efficiently capture the target spatial information from the image in a sea cucumber capture task. We segment the intended task into visual servoing phase and decision phase, modeled by LSTM layers, where the feature net share the weights with the eye gaze model. The trained net successfully automate the task in both training and unseen environment, even with disturbances. The proposed framework provides an alternative solution to underwater control tasks such as automatic capture of sea cucumbers. The automation of routined ROV tasks can potentially lower mental burden and fatigue of the operators, thus reduce human-induced errors and improve task efficiency and safety.

Our work can be improved in several directions in the future. Our framework was evaluated by an underwater object reaching and capture task, where the object stays unmooved. We plan to extend the framework and apply it to visual servo control tasks with dynamic objects and dexterous manipulation tasks such as valve opening. Moreover, such a framework needs to deploy to the real ROV platform to evaluate its feasibility in real-world experiments and its generalization capability. Thirdly, the gaze information may also encode human intentions implicitly and can be utilized in task segmentation.

ACKNOWLEDGMENT

The authors would like to thank Mr. Kunchangtai Liang for designing the calibration procedure of the eye tracker and for helping with coding.

REFERENCES

[1] L. Ji-Yong, Z. Hao, H. Hai, Y. Xu, W. Zhaoliang, and W. Lei, "Design and vision based autonomous capture of sea organism with absorptive type remotely operated vehicle," *IEEE Access*, vol. 6, pp. 73 871–73 884, 2018.

[2] S. C. Martin and L. L. Whitcomb, "Nonlinear model-based tracking control of underwater vehicles with three degree-of-freedom fully coupled dynamical plant models: Theory and experimental evaluation," *IEEE Transactions on Control Systems Technology*, vol. 26, no. 2, pp. 404–414, 2017.

[3] Z. Chu, B. Sun, D. Zhu, M. Zhang, and C. Luo, "Motion control of unmanned underwater vehicles via deep imitation reinforcement learning algorithm," *IET Intelligent Transport Systems*, 2020.

[4] Z. Chu, D. Zhu, and S. X. Yang, "Observer-based adaptive neural network trajectory tracking control for remotely operated vehicle," *IEEE Transactions on Neural networks and learning systems*, vol. 28, no. 7, pp. 1633–1645, 2016.

[5] B. Joshi and S. Rahman, "Experimental comparison of open source visual-inertial-based state estimation algorithms in the underwater domain," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[6] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[7] Y. Chen, C. Liu, L. Tai, M. Liu, and B. E. Shi, "Gaze training by modulated dropout improves imitation learning," *arXiv preprint arXiv:1904.08377*, 2019.

[8] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, and P. Abbeel, "Deep spatial autoencoders for visuomotor learning," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 512–519.

[9] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–8.

[10] H. Wu, S. Song, Y. Hsu, K. You, and C. Wu, "End-to-end sensorimotor control problems of auvs with deep reinforcement learning," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5869–5874.

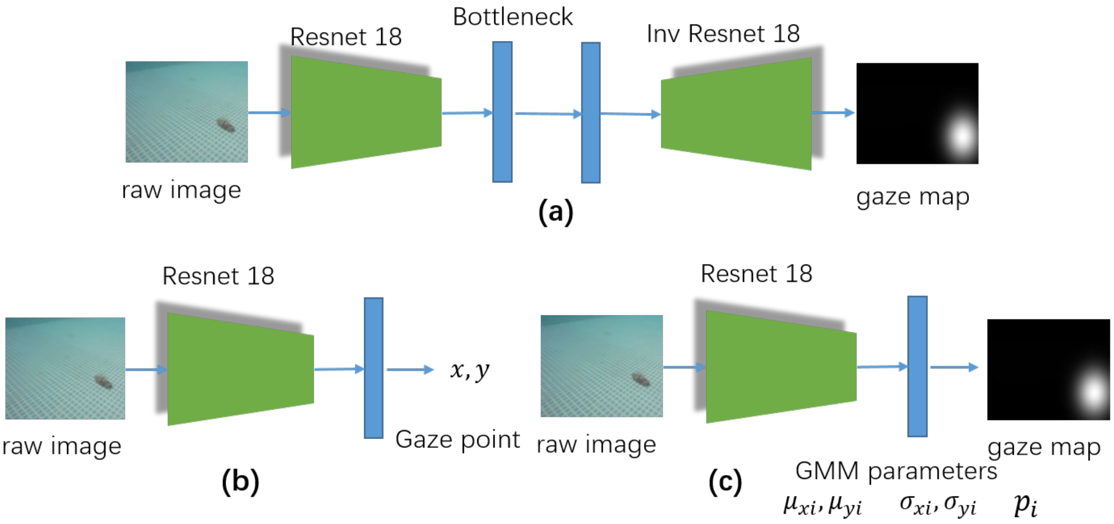

Fig. 5. Comparison of different eye gaze models. **(a)**: the model utilizing the autoencoder structure and the output target is the gaze map; **(b)**: the model whose target is the relative gaze point position in the figure; **(c)**: the model that outputs parameters for a multivariate Gaussian mixture model, which can be reconstructed into the gaze map.

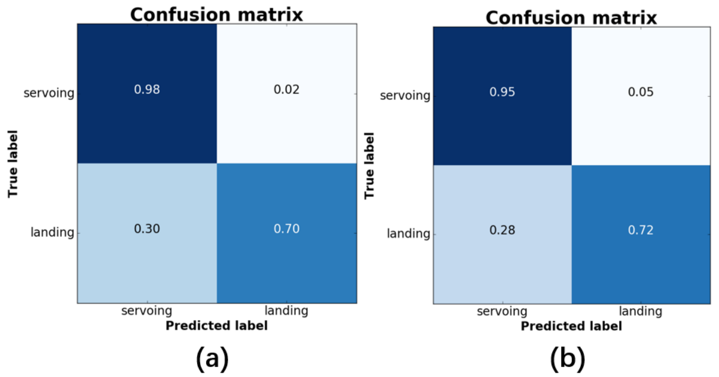

Fig. 6. Comparison of the confusion matrix for decision making models: one utilizes eye gaze information as in the proposed structure (a) and the other model that employs a spatial softmax layer as the encoder (b).

[11] M. Zhang, J. Feng, J. H. Lim, Q. Zhao, and G. Kreiman, "What am i searching for?" *ArXiv*, vol. abs/1807.11926, 2018.

[12] G.-Z. Yang, G. P. Mylonas, K.-W. Kwok, and A. Chung, "Perceptual docking for robotic control," in *International Workshop on Medical Imaging and Virtual Reality*. Springer, 2008, pp. 21–30.

[13] C. Liu, Y. Chen, L. Tai, H. Ye, M. Liu, and B. E. Shi, "A gaze

[17] M. Xu, L. Cui, P. Lv, X. Jiang, J. Niu, B. Zhou, and M. Wang, "Mdssd:

model improves autonomous driving," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. ACM, 2019, p. 33.

[14] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters *et al.*, "An algorithmic perspective on imitation learning," *Foundations and Trends® in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[15] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3216–3223.

[16] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," *arXiv preprint arXiv:1902.07296*, 2019.

Multi-scale deconvolutional single shot detector for small objects," *arXiv preprint arXiv:1805.07009*, 2018.

[18] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *arXiv preprint arXiv:1512.09300*, 2015.

[19] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1–8.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[21] R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine, "Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3758–3765.

[22] F. Provost, "Machine learning from imbalanced data sets 101," in *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, vol. 68, no. 2000. AAAI Press, 2000, pp. 1–3.