2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Enhancing Robot Perception using Vision-Aware Cognition Model

Jia Qu[1], Ryo Hanai[2], Ixchel G. Ramirez-Alpizar[2], Yukiyasu Domae[2] and Shotaro Miwa[1,2]

Abstract— **In the field of robotics, the construction of advanced perception models is essential for the successful execution of complex tasks. Traditional perception models, often grounded in cognitive frameworks, fall short in adequately processing and interpreting visual data. There is a pressing need to enhance these models with advanced visual processing capabilities. The integration of sophisticated vision models with cognitive frameworks is expected to significantly enhance the performance of perception models, yet such integrations remain underexplored.**

In this paper, we propose a novel Vision-Aware Cognition Model that effectively merges visual and cognitive components to advance robot perception. Our model integrates a cognition model, which incorporates contextual memory for nuanced long-term memory and context comprehension, with a vision model that employs spatial attention to focus on key regions of visual input. This harmonious integration enables not only robust feature extraction but also heightened adaptability to visual environmental changes.

We evaluated our model using a simulated robotic hand on a valve-turning manipulation task. By leveraging saliency visualization, we made the robot's decision-making process transparent, showcasing the distinct functions of the visual and cognitive components. The vision model demonstrates superior object segmentation, while the cognition model is adept at operation points tracking. By leveraging the strengths of both components, the proposed model achieves efficient hybrid feature extraction. Furthermore, we conducted quantitative evaluations of the model's adaptability to various visual changes, which revealed statistically significant performance improvements, highlighting its remarkable capacity for enhancing robot perception.

I. INTRODUCTION

In recent years, the field of robotics has seen remarkable progress, with its applications steadily expanding from industrial automation to medical robotics [1], [2], [3]. The automation of tasks such as robot mobility, object transportation, and manipulation calls for an enhancement in the adaptability and versatility of robotic systems in complex real-world environments. To address this challenge, the construction of advanced perception models capable of processing environmental data in ways similar to human cognitive functions is important[4], [5], [6].

Current robot perception models often adopt cognition models that aim to mimic human cognitive processes, enhancing robots' abilities to understand their environment and improve their response capabilities. These cognition models are complex systems that typically include advanced memory processing functions [7], [8] and symbolic representation [9], [10]. However, despite their proficiency in simulating certain cognitive functions, they are not sufficiently equipped to effectively process and interpret visual data using the most advanced vision processing technologies. Considering that robots largely rely on visual input for information, it is essential to augment the vision processing capabilities within these cognition models. Integrating vision models with demonstrated effectiveness in the computer vision domain, encompassing general vision functionalities like object recognition and segmentation [11], [12], [13], as well as task-oriented approaches such as transformers and spatial attention [14], [15], is expected to significantly enhance the performance of perception models.

In this paper, we propose a new Vision-Aware Cognition Model (VACM) that enhances the vision model components within a cognition model to improve robots' perception (See Fig. 1). Building upon the foundation of the cognition model that leverages contextual memory[8], our proposed model integrates this approach with a robust vision model that incorporates spatial attention mechanisms. The proposed model operates within the reinforcement learning framework and is designed to facilitate advanced robot learning. It employs the base cognition model that uses long-term memory and a deep contextual understanding to predict and sustain future actions, ensuring a continuity of behavior patterns over time. The vision model component, with its spatial attention mechanisms, empowers robots to focus on critical areas within the visual input. This integration not only enables precise feature extraction but also maintains continuous behavior patterns, thus significantly enhancing the robot's adaptability to changes in the visual environment and advancing the development of sophisticated perception models.

The primary contributions of this research are as follows:

- We propose a VACM approach that enhances the visual components within a cognition model to improve robot perception.
- By utilizing saliency visualization, we make the robot's decision-making process transparent, clarifying the collaboration between the vision and cognition model.
- We achieve a robust hybrid feature extraction by merging the precise object segmentation capabilities of the vision model with the behavior pattern learning faculties of the cognition model.
- We evaluate the effectiveness of the VACM in environments with diverse visual changes and confirm statistically significant enhancements in performance.

*This work was not supported by any organization

[1]J. Qu and S. Miwa are with the Advanced Technology R&D Center, Mitsubishi Electric Corporation, Hyogo, Japan kyoku.ka@dc.MitsubishiElectric.co.jp

[2]R. Hanai, I. Ramirez-Alpizar, and Y. Domae are with the Industrial Cyber-Physical Systems Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan domae.yukiyasu@aist.go.jp

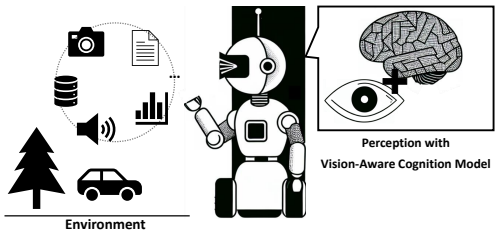


Fig. 1: Main concept: A Vision-Aware Cognition Model that enhances cognition models with advanced visual processing to augment robot perception in complex environments.

II. Related Work

A. Cognition Models in Robotics

In the field of robotics, many cognition architectures have been introduced [16]. ArmarX [17] employs a three-layered structure that simplifies software development, yet realizing such architectures can be an engineering feat, representing a substantial hurdle in practical applications.Similarly, Neuro-Serket [18] integrates cognitive modules using deep Probabilistic Graphical Models (PGMs) and facilitates model development through weakly connected modules. Despite its flexibility, the high-dimensional design space of Neuro-Serket can be difficult to manage. The WBA-PGM [19], inspired by brain research, attempts to mitigate this complexity by constraining module functions and connections, thereby streamlining the cognition model design process.

Despite these developments, the substantial engineering efforts required for implementation underscore the need for more accessible architectures. However, the implementation of these models still requires considerable engineering efforts, highlighting the need for architectures that simplify development. To this end, Reinforcement Learning (RL)-based cognition models have gained traction due to their relative ease of implementation. These models leverage memory and attention mechanisms to enhance the robotic cognition.

External memory-based approaches [20], [21] have been developed to create map-like memories from observations, but they often suffer from domain-specific limitations and lack versatility. The introduction of scene memory transformers [22] has merged memory with transformer models to address the vanishing gradient problem and maintain separate observations in memory. Nonetheless, the computational complexity and stability of transformer models in RL remain concerns.

In light of these challenges, contextual memory-based model [8] has emerged as a promising alternative. This model is not only straightforward to implement and interpret, but also demonstrates superior performance in learning behavioral patterns and maintaining stability. Given their efficacy and simplicity, contextual memory-based model is deemed suitable for our visual enhancement cognition model, offering a robust solution to the challenges faced by existing memory and attention-based methods.

B. Vision Models

Vision models have become an integral part of modern technology, with applications spanning from image classification to complex scene understanding. This section will discuss general approaches and their specialized adaptations for robotics.

1) General Vision Models: Initially, this part will explore the general vision models that have been developed through supervised learning [23]. These models form the foundation for various high-level visual tasks such as image classification [11], object detection [12] and image segmentation [13]. Image classification and object detection models, particularly those utilizing deep Convolutional Neural Networks (CNNs) [24], have been trained on extensive datasets like ImageNet [25] to accurately identify a wide range of objects. This has been a significant breakthrough for applications requiring precise visual identification, such as content filtering and medical imaging.

2) Vision Models Adapted for Robotics: In the context of robotics, vision models, particularly those based on Reinforcement Learning (RL), have been proposed to meet the demands of real-time processing and environmental interaction required in robotics.

Jeong et al. [26] devised a self-supervised image representation encoder, allowing robots to independently create reward signals, thereby enhancing task performance in real-world scenarios. Manuelli et al. [27] delved into self-supervised visual correspondence within model-based RL, establishing that predictive models markedly advance vision-dependent RL tasks. Additionally, spatial attention-based methods have been developed. Mayo et al. [14] focused on the use of spatial attention for visual navigation, implementing a convolutional net that encodes objects semantically and spatially, leading to more accurate navigation towards specific objects in a given environment. Itaya et al. [15] introduced a spatial attention mechanism into an actor-critic-based Deep RL methodology. This innovation enables mask-attention visualization, which aids in the interpretation of an agent's decision-making process from both policy and state-value perspectives. The inclusion of RL-based models with spatial attention presents a significant advantage in the field of robotics, where the efficiency of computational resources is paramount, and the necessity for prompt and precise decision-making is critical. These advancements suggest that spatial attention mechanisms are not only beneficial but are also an appropriate approach for the proposed system.

III. Method

The objective of this work is to build a model that capitalizes on the strengths of both vision and cognition model to enhance robotic perception. This section details our architecture design and explains the workings of the cognition model and the vision model respectively.

A. Vision-Aware Cognition Model Architecture

The architecture of the VACM integrates a spatial attention based vision model and a contextual memory-based cogni-

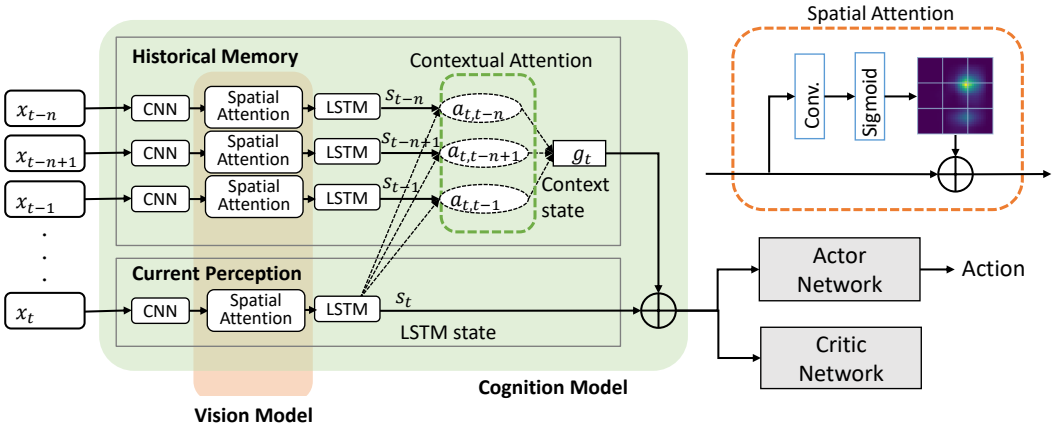


Fig. 2: Overview of the VACM. The VACM integrates a spatial attention-based vision model with a contextual memory-based cognition model, utilizing a SAC framework for reinforcement learning. Historical memory is processed through CNN layers followed by spatial attention and LSTM units to form explicit memories. Current perception is similarly treated with CNN and spatial attention, followed by LSTM for immediate context. A contextual attention mechanism refines the LSTM outputs, contributing to the context state used by both the actor and critic networks for policy training.

tion model with the base framework of Soft-Actor Critic (SAC) [28]. The details of this architect can be seen in Fig. 2.

Our cognition model is inspired by the contextual memory-based approach, initially validated in natural language processing as demonstrated by Bahdanau et al. [29], and has also shown promising performance in RL contexts [8]. This contextual memory model differentiates current perceptions from historical memory by employing cognitive strategies that mimic human-like processing for distinguishing between present experiences and past recollections. In historical memory, it stores implicit memories from LSTM and further distills these implicit memories from each timestep into explicit memory using contextual attention.

By utilizing the cognition model, we benefit from enhanced predictive capabilities and behavioral consistency. In essence, the model's integration of long-term memory and contextual attention allows for the informed and consistent action selection, ensuring a continuity of behavior patterns.

For the vision model, we integrate spatial attention mechanisms into the feature map initially extracted by a CNN layer. By applying a spatial attention mask, the feature map is selectively enhanced, providing spatial information and attentional focus for policy learning.

By the vision model, the robot can concentrate on the most pertinent aspects of the visual input, enhancing its ability to discern and react to critical elements in the environment.

It is important to note that our proposed method design is architected to be algorithm-agnostic, capable of adapting across a diverse range of RL algorithms. Our choice of SAC as the underlying algorithm was informed by its robustness and versatility, as highlighted in a comparative analysis with other fundamental RL algorithms—DQN [30] and PPO [31]—conducted by Mock et al. (2023) [32].

B. Cognition Model

The model is composed of two main parts: historical memory and current observation. The historical memory is a repository of LSTM-encoded states, each encapsulating a snapshot of past observations across a fixed temporal sequence. These stored LSTM states represent implicit memories, encoding the essence of past experiences. An attention mechanism transforms these LSTM states into an explicit context state. This process involves a weighted summarization of historical information, wherein the attention layer selectively emphasizes the most relevant stored states. By doing so, it constructs a distilled representation that reflects the temporal evolution of the environment. In parallel with the historical memory, the current observation component focuses on integrating the immediate sensory inputs. At each discrete timestep, it captures the present state of the system through a fresh LSTM encoding, providing a current perspective that supplements the historical context.

The dynamic interplay between these two units unfolds as a continual progression of the memory window, advancing from $t-n$ to $t-1$ with each timestep.

To clarify the execution at each timestep t: Firstly, the CNN generates a representation map r_t for the current observation r_t. Using this representation map r_t and the former LSTM state s_{t-1}, the LSTM then outputs the current state s_t. Then, the trajectory memory holds a historic LSTM state sequence from s_{t-1} to s_t, converting them into a context state through the application of content-based attention proposed in [29]. Context state g is defined as a weighted sum:

$$g = \sum_t \alpha_t s_t \tag{1}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{i=t-n}^{t} \exp(e_i)} \tag{2}$$

(a) (b) (c)

Fig. 3: Experiment setup. (a) ROBEL D'Claw simulation environment. (b) Observation in simulation. A black print marker distinguishes three sections of the valve and sets the rotation default in this task. (c) Action space.

$$e_t = w^T \tanh(W_\alpha s_t + b_\alpha) \tag{3}$$

Here, W_a, w, $b_\alpha$are the model parameters.

C. Vision Model

We employ a form of vision model known as spatial attention, which integrates well with the RL algorithm in our methodology. This unique strategy has been evaluated and proven to deliver high performance in Atari games [33] setup according to Itaya et al. [15]. Our attention mechanism leverages spatial attention to modulate feature maps, allowing us to emphasize those regions that are crucial for determining the optimal action and state value. The module accepts the feature map r_t extracted from the feature extractor and generates the mask attention $W(r_t)$ by applying a convolution layer and a sigmoid activation function. The mask processing for image features is computed as follows:

$$F(r_t) = r_t \cdot W(r_t) \tag{4}$$

In the equation above, $F(r_t)$ represents the masked spatial feature map for the representation map r_t. After this computation, these maps are fed into LSTM layers for contextual abstraction.

IV. IMPLEMENTATION

In this section, we describe the experimental setup, training procedures and the evaluation metrics. Then, we present the focus area and performance against environment change, comparing to the presented baseline.

A. Experimental Setup

1) Valve Rotation Task: The goal of the valve rotation task is to sustain a steady velocity (0.5 rad/sec) while performing valve rotation with three robotic fingers. This task was carried out in the ROBEL D'Claw simulation setting [34]. As depicted in Fig. 3a, the simulation environment, created in the MuJoCo physics simulator [35], features a robot hand with 9 degrees-of-freedom (DoF). Each finger has 3-DoF and the valve has 1-DoF. A prominent black marker on the valve delineates three sections to facilitate optimal operation during this task.

2) State Space: The state space for the robot consists of nine dimensions in total, three controllable joints for each of the three fingers. Since our model's objective is focused on recognizing the valve's position from the image feature, the joint angle of the valve remains unused. Thus, the state space dimension becomes dz = 9.

3) Action Space: The desired position of each fingertip was defined as action, residing within a one-dimensional (1D) manifold, with every point representing a position, as shown by different colored dotted lines in Fig. 3c. To ensure effective model learning, it was necessary to avoid discontinuity between 0 and 1; hence, we utilized the target positions of the controllable joints as actions, assigning du = 6. We also normalized actions to the range [0, 1].

4) Observation Space: The observation space combines a visual observation 32 x 32 RGB image and the encoder values of each finger joint (totaling 9D). By setting the image feature dimension at da = 8, the total observation dimension for the model (LSTM input) becomes dy = 17, which includes both the image feature and the encoder values of all finger joints.

5) Reward function: The reward function is to evaluate the performance of the robot hand implementing the valve rotation task with an emphasis on minimizing error distance at each step. It is defined as

$$\begin{aligned} r_t = -5\,|\Delta\theta_{\text{t, obj}}| + 10I\,(|\Delta\theta_{\text{t, obj}}| < 0.25) \\ +50I\,(|\Delta\theta_{\text{t, obj}}| < 0.1) \end{aligned} \tag{5}$$

where $\Delta\theta_{\text{t,obj}}$ is the error between the goal and the current object at time t.

B. Training

The policy for this task is trained in a simulated environment. For each episode, the total number of execution steps is designated as T = 80. To handle the sequence of past observations, we employ contextual attention with a window size of 20 timesteps. We also use experience replay to boost the efficiency of our training process. The agent's experiences are stored in a replay buffer that has a considerable capacity of 1e6. At each timestep during an episode, we randomly sample 256 mini-batches from this replay buffer. These batches serve to train both the policy network and two Q-function networks. The policy representation is executed through a neural network with a hidden layer consisting of 256 units.

To establish a baseline for comparison, we have implemented three configurations: the vanilla SAC, referred to as 'SAC,' which serves as the fundamental model; the SAC augmented with the vision component, denoted as 'Vision Model,'; and the SAC enhanced with the cognition component, labeled 'Cognition Model,'. This comparative analysis helps gauge the contribution and impact of the specific components within our proposed VACM. To ensure that the comparison remains fair and unbiased, we maintained identical training settings across all baseline policies. For a thorough and robust evaluation, we have trained a total of

10 models with random seeds for both our method and the baseline methods.

C. Evaluation Metrics

Our experiments evaluate task performance using the success rate metrics. Additionally, we visualize saliency to assess the type of representations the model learns.

1) Success Rate: The primary basis for evaluating tracking performance is through the success rate. It is defined as the complement of the normalized absolute distance of target error:

$$SuccessRate_t = 1.0 - \frac{target\_dist_t}{\pi} \tag{6}$$

$$target\_dist_t = |\mod(TE_t + \pi, 2\pi) - \pi| \tag{7}$$

$$TE_t = \theta_t^{\text{valve}*} - \theta_t^{\text{valve}} \tag{8}$$

where $\theta^{\text{valve}*}$ and θ^{valve} refer to the target valve position and the actual valve position at timestep t, respectively.

2) Saliency Analysis: A crucial step in understanding and interpreting the decision-making processes of robots is to determine what aspects of their input data are most influential in their decision-making. Saliency analysis serves this purpose by identifying and highlighting the elements within a robot's visual field that are pivotal to its attention and subsequent actions. Through saliency analysis, we aim to visualize on the inner workings of the robot's perception, providing tangible evidence of how the integration of visual and cognitive components in our VACM results in a more nuanced and effective approach to task execution.

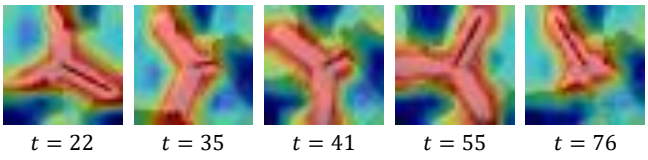
For the baseline SAC model and the SAC with cognition model, we conduct saliency analysis by adopting the perturbation-based approach detailed in [36]. This method involves the application of a localized Gaussian blur to each pixel in the input image and observing the resultant change in the policy output. By measuring the impact of these perturbations on the model's decision-making, we can generate a saliency map. This map provides a visual representation of the importance the agent places on each spatial region of the image, effectively illustrating which areas are most critical for its decision-making process. In the case of models that incorporate spatial attention mechanisms—specifically, the SAC with the vision model and our VACM—the process is inherently more straightforward. These models intrinsically produce an attention map as part of their operation, which can be directly interpreted as a saliency map. Visualizing this attention heat-map allows us to not only confirm the model's selective focus on particular features within the image but also to compare and contrast the differing attention strategies employed by each model variant.

V. Experiments and Discussion

A. Saliency Visualization

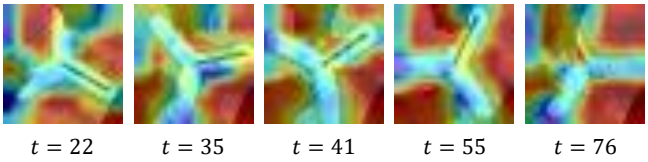
Saliency visualization has been employed to render the robot's decision-making process transparent, providing insights into the distinct functions of the visual and cognitive components. The saliency visualization results, as depicted in Fig. 4, provide a window into how our integrated VACM enhances robotic perception by merging the benefits of both models.

(a) SAC



(b) Vision Model

(c) Cognition Model

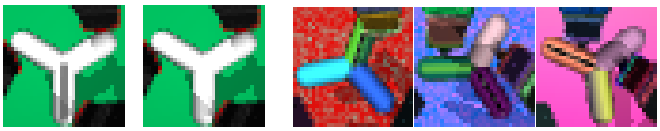


(d) VACM (Ours)

Fig. 4: Saliency visualization across different models at selected timesteps. Each subfigure (a-d) displays a sequence of images, with each image presenting a saliency map overlaid on the original observation. These heat-maps use color intensity to indicate areas of high attention, with warmer colors showing regions of maximum focus, and cooler colors indicating regions of lesser focus. The visualization demonstrates the distinct attentional strategies of each model: (a) SAC shows a focus on the marker; (b) The Vision Model displays attention to the object area; (c) The Cognition Model prioritizes operation points; and (d) VACM exhibits a hybrid focus on both object and operation points.

The baseline SAC model demonstrates a consistent focus on the print marker, indicating a reliance on rudimentary local edge features for object recognition. While this may suffice for basic object identification, the model's fixation on such elementary features becomes a critical limitation when the print marker is obscured or in the presence of visual noise, highlighting its inadequacy in dynamic and unpredictable environments.

In contrast, the SAC augmented with the vision model



(a) (b) (c)

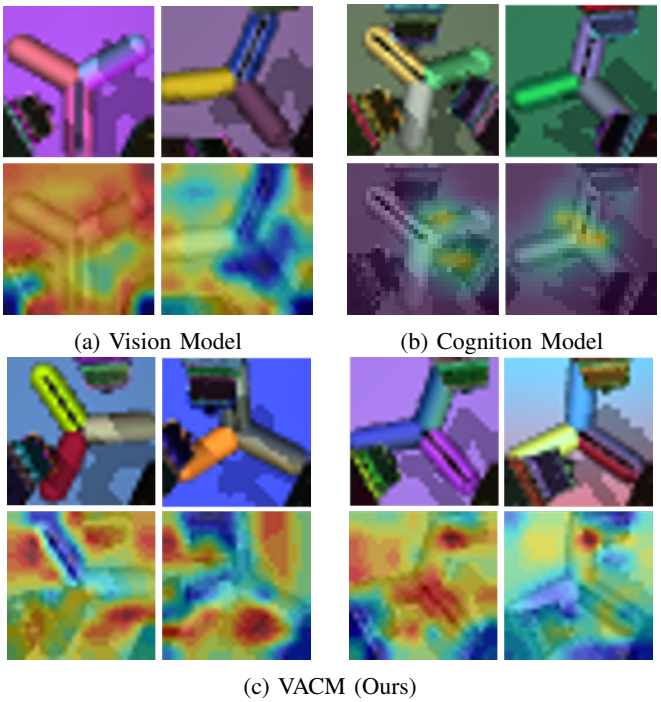
Fig. 5: Observation image samples. (a) Original observation. (b) Object change. (c) Visual change. Texture and color changes per step.

manifests a dynamic shift in focus. It begins by identifying and segmenting the valve from the background, showcasing an advanced ability to distinguish task-relevant visual features. The vision model's proficiency in discerning vital visual cues becomes evident as it adapts its focus in response to the complexities of the task environment.

On the other hand, the SAC with the cognition model directs its attention to operation points – the behavioral patterns crucial for the successful execution of the task over time. This strategic focus on operation points, rather than static object features, ensures that the robot can maintain task continuity and adapt to the evolving visual scene.

Our proposed VACM synergistically combines the vision model's finesse in feature extraction with the cognition model's strategic focus on operation points. In visual feature extraction, VACM segments background from the valve, an effect synonymous with valve object segmentation. This dual prioritization facilitates a hybrid feature extraction approach, effectively harnessing both detailed object segmentation and a profound understanding of the operationally significant patterns necessary for task completion.

Through these visualizations, we have substantiated the efficacy of our VACM in capturing the enhanced visual cues afforded by the vision model, which are crucial for complex manipulative tasks.



(a) Vision Model (b) Cognition Model

(c) VACM (Ours)

Fig. 6: A comparative performance illustration on the visual change scenario.The upper image in each panel displays observations subject to visual changes. The lower image shows the corresponding saliency maps for these observations. Saliency maps apply color gradients—warm for high-focus areas, cool for low-focus zones. The Vision Model (a) experiences difficulty when foreground and background colors blend; the Cognition Model (b) struggles with discerning operation points under challenging conditions; and our VACM (c) demonstrates adaptive focus, switching between operation point tracking and background segmentation based on the most visible feature.

B. Visual Changes Evaluation

The primary objective of this evaluation is to assess the adaptability and robustness of our proposed VACM to environmental visual changes, a critical perception ability for real-world robotic applications.

To investigate this, we conducted experiments designed to emulate realistic scenarios that a robot might encounter. The observations of each scenario are displayed in Fig. 5. The first scenario, 'Object Change,' simulates a situation where the distinctive markings or features of an object, such as a label or a color pattern, are altered or obscured. This type of change is common in real-world settings, such as when a label is covered with dirt or worn out over time. The second scenario, 'Environmental Dynamics,' represents environmental alterations where the robot must deal with variations in lighting, shadows, or background textures. This is akin to changes a robot encounters during different times or locations.

Our results, summarized in Table. I, show that while all models perform similarly in stable conditions, they react differently to these visual changes. The baseline SAC model experiences a significant performance drop, revealing its limitations in handling alterations in the visual features it was trained on.

The SAC with the vision model demonstrates greater resilience to 'Environmental Dynamics,' suggesting that its focus on feature segmentation aids in maintaining perception accuracy despite changes in the visual environment. However, its performance still diminishes slightly during the 'Object Change' scenario due to its partial dependency on specific visual markers.

The SAC with the cognition model excels in adapting to 'Object Change,' indicating its superior ability to learn and recognize operation points that are not reliant on a single visual feature. Nevertheless, it is not completely immune to the challenges posed by 'Environmental Dynamics.'

Our proposed VACM consistently delivers robust performance against both 'Object Change' and 'Environmental Dynamics.' By combining operation points tracking and background segmentation, the model demonstrates its capability

TABLE I: Main results for all experiments. Average and standard deviation of success rate over 10 models.

	Base		Object Change		Environmental Dynamics	
	Avg	Std	Avg	Std	Avg	Std
SAC	0.989	0.005	0.78	0.124	0.698	0.126
Vision Model	0.988	0.004	0.817	0.149	0.843	0.082
Cognition Model	0.981	0.006	**0.898**	0.141	0.794	0.128
VACM (Ours)	0.98	0.14	0.855	**0.112**	**0.851**	**0.045**

to handle visual perturbations. It is evident that the VACM's strengths lie not only in its superior average performance (Avg) but also in its reduced standard deviation (Std) when subjected to changes. This low standard deviation signifies a consistency in performance, suggesting that VACM is not only effective but also stable when dealing with dynamic visual environments.

Visual evidence of our model's robustness is presented in Fig. 6. While the SAC with the vision model has difficulty when the foreground and background colors blend, and the SAC with the cognition model struggles to discern operation points under similar conditions, our model adeptly adjusts its focus. It switches between operation point tracking and background segmentation, depending on which feature is more discernible under the current visual conditions. This dynamic approach significantly enhances the robot's perception ability.

VI. CONCLUSIONS

In this study, we have presented a novel VACM method aimed at enhancing robotic perception by integrating advanced vision processing capabilities with a cognitive framework. Our model leverages the strengths of spatial attention mechanisms within a robust vision model, combined with the strategic pattern recognition of a cognition model, to navigate and interpret complex visual environments with a human-like understanding.

In our experiments, we have implemented saliency visualization to explore the collaborative functions of the visual and cognitive elements of our model. The results indicate that our VACM effectively melds the vision model's precision in object segmentation with the cognition model's strategic emphasis on key operation points.

Furthermore, we have quantitatively evaluated our model across a variety of environments involving visual changes that robots are expected to face in the real world. The results confirm the effectiveness of our VACM with statistically significant enhancements in performance when compared to existing models.

Our findings underscore the potential of VACM to substantially enhance robotic perception. The model's demonstrated resilience to a wide range of visual variations greatly augments the reliability and operational efficiency of autonomous robotic systems. Future research will aim to apply this vision-aware cognition model to more complex robotic tasks over a long horizon.

REFERENCES

[1] B. C. Kok and H. Soh, "Trust in robots: Challenges and opportunities," *Current Robotics Reports*, vol. 1, no. 4, pp. 297–309, 2020.

[2] J. Arents and M. Greitans, "Smart industrial robot control trends, challenges and opportunities within manufacturing," *Applied Sciences*, vol. 12, no. 2, p. 937, 2022.

[3] P. E. Dupont, B. J. Nelson, M. Goldfarb, B. Hannaford, A. Menciassi, M. K. O'Malley, N. Simaan, P. Valdastri, and G.-Z. Yang, "A decade retrospective of medical robotics research from 2010 to 2020," *Science robotics*, vol. 6, no. 60, p. eabi8017, 2021.

[4] C. Premebida, R. Ambrus, and Z. Marton, *Intelligent Robotic Perception Systems*, 11 2018.

[5] H. Yan, M. H. Ang, and A. N. Poo, "A survey on perception methods for human–robot interaction in social robots," *International Journal of Social Robotics*, vol. 6, pp. 85–119, 2014.

[6] M. Pfeiffer, M. Schaeuble, J. Nieto, R. Siegwart, and C. Cadena, "From perception to decision: A data-driven approach to end-to-end motion planning for autonomous ground robots," in *2017 ieee international conference on robotics and automation (icra)*. IEEE, 2017, pp. 1527–1533.

[7] M. A. Goodrich, "Using models of cognition in hri evaluation and design." in *AAAI Technical Report (5)*, 2004, pp. 17–24.

[8] J. Qu, S. Miwa, and Y. Domae, "Interpretable navigation agents using attention-augmented memory," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2022, pp. 2575–2582.

[9] S. Lemaignan, R. Ros, E. A. Sisbot, R. Alami, and M. Beetz, "Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction," *International Journal of Social Robotics*, vol. 4, pp. 181–199, 2012.

[10] V. M. De La Cruz, A. Di Nuovo, S. Di Nuovo, and A. Cangelosi, "Making fingers and words count in a cognitive robot," *Frontiers in behavioral neuroscience*, vol. 8, p. 13, 2014.

[11] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[12] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[13] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[14] B. Mayo, T. Hazan, and A. Tal, "Visual navigation with spatial attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 898–16 907.

[15] H. Itaya, T. Hirakawa, T. Yamashita, H. Fujiyoshi, and K. Sugiura, "Visual explanation using attention mechanism in actor-critic-based deep reinforcement learning," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–10.

[16] T. Taniguchi, S. Murata, M. Suzuki, D. Ognibene, P. Lanillos, E. Ugur, L. Jamone, T. Nakamura, A. Ciria, B. Lara, *et al.*, "World models and predictive coding for cognitive and developmental robotics: Frontiers and challenges," *Advanced Robotics*, vol. 37, no. 13, pp. 780–806, 2023.

[17] N. Vahrenkamp, M. Wächter, M. Kröhnert, K. Welke, and T. Asfour, "The robot software framework armarx," *it-Information Technology*, vol. 57, no. 2, pp. 99–111, 2015.

[18] T. Taniguchi, T. Nakamura, M. Suzuki, R. Kuniyasu, K. Hayashi, A. Taniguchi, T. Horii, and T. Nagai, "Neuro-serket: development of integrative cognitive system through the composition of deep

probabilistic generative models," *New Generation Computing*, vol. 38, pp. 23–48, 2020.
[19] T. Taniguchi, H. Yamakawa, T. Nagai, K. Doya, M. Sakagami, M. Suzuki, T. Nakamura, and A. Taniguchi, "A whole brain probabilistic generative model: Toward realizing cognitive architectures for developmental robots," *Neural Networks*, vol. 150, pp. 293–312, 2022.
[20] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2616–2625.
[21] J. F. Henriques and A. Vedaldi, "Mapnet: An allocentric spatial memory for mapping environments," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8476–8484.
[22] K. Fang, A. Toshev, L. Fei-Fei, and S. Savarese, "Scene memory transformer for embodied agents in long-horizon tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 538–547.
[23] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.
[24] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.
[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
[26] R. Jeong, Y. Aytar, D. Khosid, Y. Zhou, J. Kay, T. Lampe, K. Bousmalis, and F. Nori, "Self-supervised sim-to-real adaptation for visual robotic manipulation," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 2718–2724.
[27] L. Manuelli, Y. Li, P. Florence, and R. Tedrake, "Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning," *arXiv preprint arXiv:2009.05085*, 2020.
[28] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.
[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
[30] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.
[31] L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, *et al.*, "Model-based reinforcement learning for atari," *arXiv preprint arXiv:1903.00374*, 2019.
[32] J. W. Mock and S. S. Muknahallipatna, "A comparison of ppo, td3 and sac reinforcement algorithms for quadruped walking gait generation," *Journal of Intelligent Learning Systems and Applications*, vol. 15, no. 1, pp. 36–56, 2023.
[33] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.
[34] M. Ahn, H. Zhu, K. Hartikainen, H. Ponte, A. Gupta, S. Levine, and V. Kumar, "Robel: Robotics benchmarks for learning with low-cost robots," in *Conference on robot learning*. PMLR, 2020, pp. 1300–1313.
[35] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.
[36] S. Greydanus, A. Koul, J. Dodge, and A. Fern, "Visualizing and understanding atari agents," in *International conference on machine learning*. PMLR, 2018, pp. 1792–1801.