2023 IEEE Conference on Artificial Intelligence (CAI) | 979-8-3503-3984-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAI54212.2023.00035

Robust Active Simultaneous Localization and Mapping Based on Bayesian Actor-Critic Reinforcement Learning

Bryan Pedraza and Dimah Dera
Department of Electrical and Computer Engineering
The University of Texas Rio Grande Valley
bryan.pedraza01@utrgv.edu, dimah.dera@utrgv.edu

***Abstract*— Autonomous mobile robots play vital roles in business, industry, manufacturing, e-commerce, and healthcare. Autonomous navigation and obstacle avoidance involve localizing a robot to actively explore and map an unknown environment autonomously without prior knowledge. Simultaneous localization and mapping (SLAM) present a severe challenge. This paper proposes a novel approach for robust navigation and robot action mapping based on Bayesian Actor-Critic (A2C) reinforcement learning. The principle of Actor-Critic combines policy-based and value-based learning by splitting the model into two: the policy model (Actor) computes the action based on the state, and the value model (Critic) tracks whether the agent is ahead or behind during the game. That feedback guides the training process, where both models participate in a game and optimize their output as time passes. We develop a Bayesian A2C model that generates robot actions and quantifies uncertainty on the actions toward robust exploration and collision-free navigation. We adopt the Bayesian inference and optimize the variational posterior distribution over the unknown model parameters using the evidence lower bound (ELBO) objective. The first-order Taylor series approximates the mean and covariance of the variational distribution passed through non-linear functions in the A2C model. The propagated covariance estimates the robot's action uncertainty at the output of the Actor-network. Experiments demonstrate the superior robustness of the proposed Bayesian A2C model exploring heavily noisy environments compared to deterministic homologs. The proposed framework can be applied to other fields of research (underwater robots, biomedical devices/robots, micro-robots, drones, etc.) where robustness and uncertainty quantification are critical.**

Keywords*— Simultaneous localization and mapping (SLAM), uncertainty, Actor-Critic (A2C), variational inference (VI), and evidence lower bound (ELBO).

I. Introduction

Reinforcement learning (RL) has been an area of great interest in machine learning for quite some time due to the remarkable performance where agents could potentially learn and outperform human-level skills in tasks related environments [1]-[5]. RL has achieved tremendous success in several research areas; one area of interest is active simultaneous localization and mapping (SLAM). Active SLAM is a set of approaches in which an autonomous robot localizes itself and simultaneously navigates and maps an unknown environment [5]-[9].

The main purpose of active SLAM is to efficiently explore unknown/unseen environments and gather valuable data to construct a map of the real-world environment. SLAM technology has demonstrated great potential in many applications, such as autonomous mobile robotics, autonomous vehicles with two lasers, LIDAR (light imaging detection and ranging) and RADAR (radio detection and ranging), augmented reality, surveillance with unmanned air vehicles, and autonomous drones [10]-[13]. Many approaches were introduced in the literature for solving active SLAM problems, such as filter-base, graph-based, and optimization-based methods [14]-[16]. However, exploring unknown environments remains challenging due to a lack of prior knowledge and the critical need for quantifying uncertainty toward robust exploration in heavily noisy environments.

Researchers have been continuously developing new techniques, methods, and solutions to ultimately improve the performance and robustness of active SLAM. RL approaches, such as Q-learning, Deep Q-Networks (DQN), double DQN, dueling double DQN, Actor-Critic (A2C) and its variations have shown promising results [17]-[21]. Nevertheless, the uncertainty in a predicted action and its relation to robot behavior in noisy environments have not been investigated in the literature. Current robot systems lack resilient mechanisms for exploration and uncertainty mitigation, which are critical for intelligent decision-making.

This paper proposes a novel Active SLAM algorithm based on Bayesian Actor-Critic (A2C) deep reinforcement learning model. The proposed Bayesian A2C estimates the robot's actions and quantifies the uncertainty in the actions. The uncertainty can guide the robot to move toward the direction of low uncertainty and avoid collisions. We adopt Bayesian inference and optimize the variational posterior distribution over the A2C model parameters given the data. Thus, the optimization is achieved by minimizing the Kullback-Leibler (KL) divergence between the approximate and true posterior distributions. The objective function is known as the evidence lower bound (ELBO). We extend the uncertainty propagation framework in [22] by propagating the mean and covariance of the variational posterior through the Bayesian A2C model layers. The first-order Taylor linearization is used to approximate the mean and covariance after non-linear layers.

979-8-3503-3984-0/23/$31.00 ©2023 IEEE
DOI 10.1109/CAI54212.2023.00035

The propagated covariance at the output of the Actor network implies the uncertainty over the predicted action. Quantifying uncertainty over the robot's actions improves the robustness of the autonomous navigation of the robot while exploring noisy environments. Our experiments show that the Bayesian A2C model obtains higher reward values as compared to the deterministic A2C model. Moreover, Bayesian A2C maintains its performance (similar reward values) when the robot explores a noisy environment (corrupted with random noise) during validation (without being trained on the noisy environment).

This study investigates the fundamental issues of intelligent robotic systems' robustness and trustworthiness by developing a principled theoretical and algorithmic framework that impacts the design of mapping and active sensing algorithms for autonomous robot systems. Such capabilities are crucial in security, surveillance, and environmental monitoring, where it is necessary to quickly gain situational awareness of the terrain, buildings, and humans in the environment.

II. Bayesian Actor-Critic Reinforcement Learning

The actor-critic model is a temporal difference (TD) learning method that represents the policy function independent of the value function. The policy function returns the agent's actions based on a given state. At the same time, the value function determines the expected return for an agent starting at a given state and acting according to a particular policy. In the Actor-Critic method, the policy is referred to as the *actor* that proposes a set of possible actions given a state, and the estimated value function is referred to as the *critic*, which evaluates actions taken by the actor based on the given policy. We consider convolution neural networks (CNNs) in both the actor and critic models and introduce a prior distribution over the networks' weights $\boldsymbol{\Omega}\sim p(\boldsymbol{\Omega})$. The variational posterior distribution of the weights given the data is $p(\boldsymbol{\Omega}|\boldsymbol{\mathcal{D}}_t)$, where $\boldsymbol{\mathcal{D}}_t = \{\boldsymbol{s}_t, \boldsymbol{a}_t\}$, and $\boldsymbol{s}_t, \boldsymbol{a}_t$ are the state and action at time t. The variational posterior $p(\boldsymbol{\Omega}|\boldsymbol{\mathcal{D}}_t)$ is approximated using the variational inference by optimizing the ELBO objective $\mathcal{L}(\boldsymbol{\phi}; \boldsymbol{\mathcal{D}}_t)$ in (1).

$$\mathcal{L}(\boldsymbol{\phi}; \boldsymbol{\mathcal{D}}_t) = E_{q_{\boldsymbol{\phi}}(\boldsymbol{\Omega})}\{\log p(\boldsymbol{\mathcal{D}}_t\,|\boldsymbol{\Omega})\} - KL\big[q_{\boldsymbol{\phi}}(\boldsymbol{\Omega})\,||p(\boldsymbol{\Omega})\big], \quad (1)$$

where $\boldsymbol{\phi}$ represents the variational parameters, i.e., the mean and covariance of the variational posterior, $q_{\boldsymbol{\phi}}(\boldsymbol{\Omega})$ and $t = 1, \ldots, T$, is the number of training steps per epoch.

The ELBO objective (or loss) function $\mathcal{L}(\boldsymbol{\phi}; \boldsymbol{\mathcal{D}}_t)$ consists of two parts: the negative expected log-likelihood of the training data (environment maps) given the network parameters and a regularization term. The latter is defined by the KL divergence between the proposed variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{\Omega})$ and the prior distribution $p(\boldsymbol{\Omega})$. The ELBO loss function is minimized as part of the total loss function using the gradient descent update rule during training of the Bayesian A2C networks. The total objective function of the Bayesian A2C model is given as,

$$\mathrm{J}(\phi, \theta) = \sum_{t=0}^{T} \mathcal{L}(\phi; \boldsymbol{\mathcal{D}}_t)\, A(\boldsymbol{s}_t, \boldsymbol{a}_t), \quad (2)$$

$$A(\boldsymbol{s}_t, \boldsymbol{a}_t) = R(\boldsymbol{s}_t, \boldsymbol{a}_t) + V_\theta(\boldsymbol{s}_{t+1}) - V_\theta(\boldsymbol{s}_t), \quad (3)$$

where $A(\boldsymbol{s}_t, \boldsymbol{a}_t)$ is the TD loss, R is the reward, $V_\theta(\boldsymbol{s}_{t+1})$ is the average value of the state, and θ is the parameters of the Critic network. Figure 1 shows a schematic description of the Bayesian A2C model, including the Actor and Critic networks.

We adopt the first-order Taylor approximation to propagate the mean and covariance matrix after non-linear activation functions in the model. Let $\mathbf{z} = f(\mathbf{x})$, where $\mathbf{x}$ and $\mathbf{z}$ are the feature maps before and after activation function, f, for a layer in the network (Fig. 2). The mean and covariance of $\mathbf{x}$ propagate through f to $\mathbf{z}$ using the first-order approximation as follows,

$$\boldsymbol{\mu}_\mathbf{z} \approx f(\boldsymbol{\mu}_\mathbf{x}),$$
$$\boldsymbol{\Sigma}_\mathbf{z} \approx \boldsymbol{\Sigma}_\mathbf{x} \odot (\nabla f(\boldsymbol{\mu}_\mathbf{x}) \nabla f(\boldsymbol{\mu}_\mathbf{x})^T). \quad (4)$$

The derivation of the mean and covariance applies to any non-linear function in the model. By propagating the mean and covariance of the variation distribution, we obtain the predictive distribution over any new action $\hat{\boldsymbol{a}}$, given a state $\hat{\boldsymbol{s}}$ and the training data $\boldsymbol{\mathcal{D}}$, i.e., $p(\boldsymbol{a}_t|\boldsymbol{s}_t, \boldsymbol{\mathcal{D}})$, by marginalizing out the model parameters $\boldsymbol{\Omega}$. The mean of $p(\boldsymbol{a}_t|\boldsymbol{s}_t, \boldsymbol{\mathcal{D}})$ represents the robot's action (step or move), while the covariance matrix reflects the uncertainty associated with the predicted action.

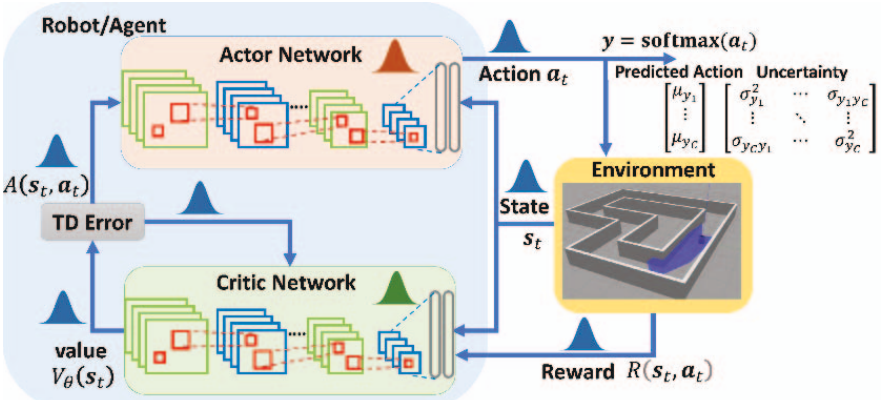


Fig 1. A schematic description of the proposed Bayesian A2C model with the two convolution neural networks: Actor network and Critic network. The parameters of the two networks are random variables following variational posterior distribution. We propagate the mean and covariance of the variational distribution through the non-linear layers of the two networks. The output of the Actor network is the predictive distribution over the action, $p(\boldsymbol{a}_t|\boldsymbol{s}_t, \boldsymbol{\mathcal{D}})$. The mean and covariance of $p(\boldsymbol{a}_t|\boldsymbol{s}_t, \boldsymbol{\mathcal{D}})$ refer to the predicted robot's action and the uncertainty associated with that action.

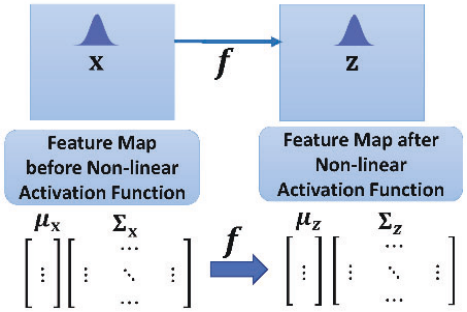


Fig 2. A schematic example of the mean-covariance propagation. The first-order Taylor approximation facilitates the propagation of the mean and covariance matrix of the variational posterior through the non-linear activation functions in the proposed Bayesian A2C networks.

The process of using Bayesian A2C model to conduct Active SLAM training for mobile robots is shown in Algorithm 1.

Algorithm 1. The proposed Bayesian Actor-Critic Reinforcement Learning.

Require: Total number of training epochs, Max-epoch, maximum number of training steps per epoch T, and learning rate η.

1. Initialize the parameters of the Actor and Critic networks.
2. Set $t = 0$, and epoch $= 0$.
3. **While** epoch < Max-epoch
4. Initialize the input environment to get the initial state $\boldsymbol{s}_0$.
5. **While** $t < T$
6. According to the observed state $\boldsymbol{s}_t$, the current Actor neural network selects the actions $\boldsymbol{a}_t$ to be executed.
7. Apply the selected action to the current environment for immediate rewards $R(\boldsymbol{s}_t, \boldsymbol{a}_t)$ and new environment status $\boldsymbol{s}_{t+1}$.
8. Calculate the value $V_\theta(\boldsymbol{s}_t)$ and the TD loss $A(\boldsymbol{s}_t, \boldsymbol{a}_t)$.
9. Calculate the loss $\mathrm{J}(\phi, \theta) = \sum_{t=0}^{T} \mathcal{L}(\phi; \mathcal{D}_t)\, A(\boldsymbol{s}_t, \boldsymbol{a}_t)$.
10. Using ADAM optimizer to optimize model parameters.
11. Update network parameters with learning rate η, $\phi \leftarrow \phi + \eta \nabla_\phi \mathrm{J}(\phi, \theta)$ and $\theta \leftarrow \theta + \eta \nabla_\theta \mathrm{J}(\phi, \theta)$
12. **end While**
13. **end While**

III. Experiments and Results

We conduct experiments using the Gazebo simulator as an open-source Open-AI gym extension. The Linux operating system and Robot Operating System (ROS) are also used to construct the software framework, libraries, messages and tools for the robot and its environment. We compare the proposed Bayesian A2C model with the deterministic (DET) A2C model. The proposed Bayesian A2C and the DET A2C models are trained, tested, and examined in a simulated Gazebo environment. We evaluate the robustness of the proposed Bayesian A2C model against noisy environments by adding random (Gaussian) noise to images/frames during testing. In real-world applications, many external factors, such as outside signals, may distort the images, thus causing noisy corrupted images. The noise analysis is performed for both the proposed Bayesian and DET models for fair comparison. The Actor and Critic CNNs are trained with the same learning rate and batch size (16). The size of the input image/frame is 32×32. We use the SoftMax function at the output of the Actor network to classify (three) actions as moving straight, right, or left. The CNNs architecture is 10 layers with 32 kernels (5×5) in the first 3 layers, 64 kernels (3×3) in the following 3 layers and 128 kernels (1×1) in the subsequent 3 layers. The last layer is fully-connected. The performance is evaluated using the moving average and the cumulative rewards.

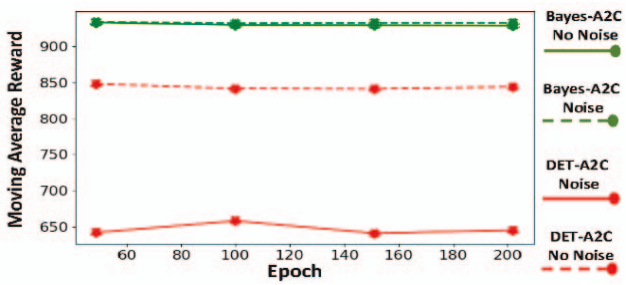
Figs. 3(a) and 3(b) demonstrate the moving average reward and the cumulative reward for the proposed Bayesian (Bayes) A2C model (green curves) and deterministic (DET) A2C model (red curves). The experiment includes two cases: a clean (No Noise) environment (dashed curves) and a noisy (or corrupted) environment (solid curves). We notice that the moving average as well as the cumulative rewards of the DET model are drastically higher in the clean environment as compared to the noisy environment. The rewards highly decrease when the environment becomes noisy (solid curve).

On the other hand, the Bayesian A2C model maintains its performance under noisy environments. The moving average and cumulative reward presented by the green curves (both dashed and solid ones) have almost the same values in clean and noisy cases. Thus, the proposed Bayesian A2C model manifests superior robustness against noisy environments, as evidenced by the rewards values.

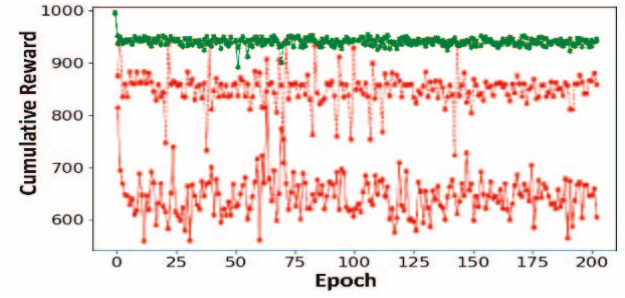
We link the robustness of the proposed Bayesian A2C model against noisy environments with the model's ability to propagate uncertainty through the Actor and Critic networks' layers and non-linearities. The uncertainty is given by the covariance matrix of the variational posterior distribution of the Actor and Critic networks parameters. We consider that during navigation and exploration of the environment (during training), the availability of additional information in the form of the second moment (covariance matrix) helps the learning process and results in the proposed model that is more robust to noise.

IV. Conclusion

In this paper, we proposed a robust and uncertainty-aware active simultaneous localization and mapping (SLAM) approach based on Bayesian Actor-Critic (A2C) reinforcement learning. We adopt Bayesian inference and propagate the mean and covariance matrix of the variational posterior distribution over the model parameters through the layers of the Actor and Critic networks. Using the first-order Taylor approximation, we estimate the mean and covariance after non-linear activation functions in the networks. Propagating the mean and covariance of the variational distribution allows computing the mean and covariance of the predictive distribution over the robot's action. The mean of the predictive distribution represents the predicted robot's action, while the covariance matrix provides information about the uncertainty in that action (or move). Our proposed models have significantly enhanced the robust navigation in noisy environments. The experimental results using the Open-AI gym Gazebo simulator have established superior robustness of the proposed model against Gaussian noise in the simulated environment as compared to the deterministic A2C model.



(a) Moving Average Reward



(b) Cumulative Reward

Fig 3. The performance evaluation of the proposed Bayesian (Bayes) A2C model as compared to the deterministic (DET) A2C model measured by (a) Moving average reward (left graph) and (b) cumulative reward (right graph). The rewards of the Bayesian A2C are presented in green color, while the rewards of the DET A2C model are presented in red color. The evaluation includes two cases: a clean (No noise) environment (dashed curves) and a noisy environment (solid curves).

In future work, we plan to investigate the robustness of the proposed Bayesian A2C model against attacked environments by introducing different types and levels of adversarial attacks during testing.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation award NSF CRII-2153413. Bryan Pedraza is supported by the US Department of Education through a Graduate Assistance in Areas of National Need (GAANN) program Award Number P200A210140-22.

REFERENCES

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves *et al*. "Human-level control through deep reinforcement learning." *Nature* 518, no. 7540, pp. 529-533, 2015.

[2] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates." In *IEEE international conference on robotics and automation (ICRA),* pp. 3389-3396, May 2017.

[3] R. Liu, F. Nageotte, P. Zanne, M. de Mathelin, and B. Dresp-Langley. "Deep reinforcement learning for the control of robotic manipulation: a focussed mini-review." *Robotics*, 10(1), p.22, 2021.

[4] M. F. Ahmed, K. Masood, and V. Fremont, "Active SLAM: A Review On Last Decade." *arXiv preprint arXiv:2212.11654*, 2022.

[5] M. Pumperla, and K. Ferguson. "Deep learning and the game of Go." Vol. 231. Shelter Island, NY, USA: Manning Publications Company, 2019.

[6] D. Trivun, E. Šalaka, D. Osmanković, J. Velagić and N. Osmić, "Active SLAM-based algorithm for autonomous exploration with mobile robot," *IEEE International Conference on Industrial Technology (ICIT),* Seville, Spain, pp. 74-79, 2015.

[7] N. Palomeras, M. Carreras, and J. Andrade-Cetto. "Active SLAM for autonomous underwater exploration." *Remote Sensing*, 11(23), p.2827, 2019

[8] Y. Chen, S. Huang and R. Fitch, "Active SLAM for Mobile Robots With Area Coverage and Obstacle Avoidance," in *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 3, pp. 1182-1192, June 2020.

[9] M. Mihálik, B. Malobický, P. Peniak, P. Vestenický. "The New Method of Active SLAM for Mapping Using LiDAR." *Electronics*, 11(7):1082, 2022.

[10] Y. Liu, M. Xu, G. Jiang, X. Tong, J. Yun, Y. Liu, B. Chen, Y. Cao, N. Sun, and Z. Li, "Target localization in local dense mapping using RGBD SLAM and object detection." *Concurrency and Computation: Practice and Experience*, 34(4), p.e6655, 2022.

[11] D. Van Nam, and K. Gon-Woo. "Solid-state LiDAR based-SLAM: A concise review and application." In *IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 302-305, 2021.

[12] C. Cadena et al., "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age," in *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309-1332, Dec. 2016, doi: 10.1109/TRO.2016.2624754.

[13] A. Gupta, and X. Fernando. "Simultaneous localization and mapping (SLAM) and data fusion in unmanned aerial vehicles: Recent advances and challenges." *Drones* 6, no. 4: 85, 2022.

[14] J. A. Placed, and J. A. Castellanos. "A General Relationship between Optimality Criteria and Connectivity Indices for Active Graph-SLAM." *IEEE Robotics and Automation Letters*, 2022.

[15] M. Kontitsis, E. Theodorou and E. Todorov, "Multi-robot active SLAM with relative entropy optimization," *American Control Conference*, 2013.

[16] A. Khairuddin, M. Talib and H. Haron, "Review on simultaneous localization and mapping (SLAM)," *IEEE International Conference on Control System, Computing and Engineering (ICCSCE),* 2015.

[17] H. Van Hasselt, A. Guez, D. Silver, "Deep reinforcement learning with double Q-learning." In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence,* pp. 2094–2100, February 2016.

[18] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas. "Dueling network architechtures for deep reinforcement learning." In *International Conference on Machine Learning*, pp. 1995-2003, June 2016.

[19] J. A. Placed, and J. A. Castellanos. "A deep reinforcement learning approach for active SLAM." *Applied Sciences*, 10(23), p.8386, 2020.

[20] S. Wen, Z. Ji, A. B. Rad, Z. Guo. "A Hybrid Technique for Active SLAM Based on RPPO Model with Transfer Learning." *Research Square*; 2022. DOI: 10.21203/rs.3.rs-1229897/v1.

[21] M. Alcalde, M. Ferreira, P. González, F. Andrade and G. Tejera, "DA-SLAM: Deep Active SLAM based on Deep Reinforcement Learning," 2022 Latin American Robotics Symposium (LARS), 2022 Brazilian Symposium on Robotics (SBR), and 2022 Workshop on Robotics in Education (WRE), São Bernardo do Campo, Brazil, 2022, pp. 282-287, doi: 10.1109/LARS/SBR/WRE56824.2022.9996006.

[22] D. Dera, N. C. Bouaynaya, G. Rasool, R. Shterenberg and H. M. Fathallah-Shaykh, "PremiUm-CNN: Propagating Uncertainty Towards Robust Convolutional Neural Networks." In *IEEE Transactions on Signal Processing*, vol. 69, pp. 4669-4684, 2021.