Structured Bayesian Meta-Learning for Data-Efficient Visual-Tactile Model Estimation

Shaoxiong Yao[1], Yifan Zhu[2], and Kris Hauser[1]

***Abstract*—Robots ought to fuse vision and touch data to predict how a deformable object reacts during manipulation in unstructured settings. Estimating such a visual-tactile model usually requires a lot of data since vision suffers from occlusion and touch data is sparse and noisy. This paper proposes a novel method, structured Bayesian meta-learning (SBML), to allow data-efficient visual-tactile model estimation for diverse and heterogeneous deformable objects. SBML uses perception to define an object structure that establishes a common parameter space for all meta-training and testing objects, regardless of size and shape. SBML was applied to the recently proposed volumetric stiffness field (VSF) visual-tactile model. Experiments show that in two classes of heterogeneous objects, namely plants and shoes, SBML outperforms existing approaches in terms of force and torque prediction accuracy in zero- and few-shot settings.**

I. Introduction

Vision alone cannot accomplish robust deformable object manipulation in unstructured settings and fusing tactile information with vision is essential for many tasks. A visual-tactile model that predicts how a deformable object reacts during manipulation enables manipulation in clutter [8], assistive dressing [5], and complaint tool usage [10]. Estimating a visual-tactile model usually requires a lot of data (e.g. 30k transitions in [10]) since vision suffers from occlusion and touch data is sparse and noisy. To learn models efficiently, past work has assumed restricted contact regions during tool usage [17, 2], homogeneity of the object [6, 14, 15], or sim2real transfer of the visual-tactile model [4, 16]. These assumptions reduce the representation power (i.e., capacity) of the visual-tactile models.

In this paper, we present a few-shot learning method to build high-capacity visual-tactile models of an object using vision data and a small number of touches. Our approach builds a *prior distribution of visual-tactile models* from visual-tactile experience from related objects, and addresses the few-shot learning problem as an estimation of a Bayesian posterior. This general approach is known as Bayesian meta-learning [7, 19]. For example, the experience that plants' leaves are usually soft while branches are usually stiff will help build accurate models of a novel plant. Moreover, the model should improve quickly from zero- to few-shots of experience on a novel plant.

Standard Bayesian meta-learning methods assume fixed dimensional input and output spaces, but visual-tactile models of different deformable objects have a different number of state variables and material parameters. For example, the finite element models of two objects have different mesh structures. We introduce a novel *structured Bayesian meta-learning* (SBML) approach that accommodates diverse structures $\mathcal{S}$ that capture the object's spatial structure as well as visual appearance. The structural components $\mathcal{S}$ connect meta-parameters from a "universal" fixed-dimension space to the primary visual-tactile model on the object's computational mesh. We then apply a hierarchical Bayes maximum likelihood estimation approach to learn the meta-parameters. The learned meta-parameters can leverage a novel object's structures to predict material parameter priors, enhancing online few-shot estimation accuracy.

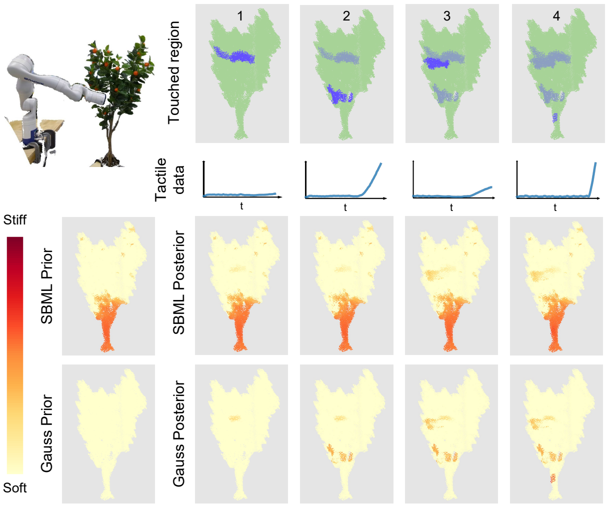

Fig. 1. Our structured Bayesian meta-learning (SBML) approach enables few-shot learning of a novel plant's force response using experience interacting with past plants. The bottom two rows compare our SBML method against a stiffness estimator using naïve uniform Gaussian prior. In the zero-shot prediction (left column), SBML already estimates the trunk and branches as stiffer than the leaves and its predictions improve as it touches the plant. Based on the touched region (top row) and tactile feedback (second row, norm of joint torques over time), the updated estimates after each touch are shown in columns 1–4. The naïve prior needs at least 10 touches to reach accuracy comparable to our method's zero-shot estimate (see Tab. I).

We use SBML to estimate structured deformable object models of heterogeneous plants and household objects. We apply the method to the recently developed volumetric stiffness field (VSF) approach [18] with tens of thousands of parameters. Our approach improves the performance of VSF above uninformed baselines and an unstructured state-of-the-art meta-learning approach and an informed prior generated by SBML can be seen qualitatively in Fig. 1. We examine SBML's performance with different meta-training datasets and

[1]: S. Yao and K. Hauser are with the Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA. {syao16, kkhauser}@illinois.edu,[2]: Y. Zhu is with the Department of Computer Science, Yale University, CT, USA. yifan.zhu@yale.edu

show its adaptability to various structural assumptions, such as foundation model features (DINOv2).

II. Method

A. Structured Bayesian meta-learning

A standard meta-learning problem has input $x \in \mathbb{R}^m$ and output $y \in \mathbb{R}^n$. We assume that each output y follows a distribution $y \sim P(Y|x, \alpha)$ conditioned on both the input variables x and an unobservable *task* α. A meta-training dataset $D = \{D_\alpha \,|\, \alpha = 1, \ldots, L\}$ has data from multiple tasks. The *online learning* task is to predict $P(y|x, \alpha^*)$ for a novel task α^* given data from a support set $D_* = \{(x_*^i, y_*^i) \,|\, i = 1 \ldots, N_*\}$, from the zero-shot ($N_* = 0$) to the few-shot (small N_*) cases.

We approximate the likelihood using a parameterized function $P(y|x, \alpha) = f(y; x, \theta)$, where $\theta \in \mathbb{R}^M$ are the learnable model parameters. Bayesian meta-learning learns a distribution of model parameters $\theta \sim P(\theta|\psi)$ from the meta-training dataset D [7, 19]. $P(\theta|\psi)$ is the *prior* of model parameters and ψ are the meta-parameters shared across tasks. Given online data, we can first find a maximum a posteriori (MAP) estimate $\hat{\theta}$ for the novel task α^* and use it to predict y from x.

Standard Bayesian meta-learning assumes the same parametric function maps x to y across all tasks. This assumption does not hold when tasks have varied input-to-output mapping structures, like visual-tactile model estimation. Here, the task α indicates the specific object being interacted with. When objects are heterogeneous and diverse, each object has a different number of parameters θ_α to estimate, i.e., $\theta_\alpha = \mathbb{R}^{M_\alpha}$.

SBML assumes that a task-specific structure $\mathcal{S}_\alpha$ is instantiated for each task α from an auxiliary *structure generation module*, outside the learning pipeline. The model output distribution is conditioned on both input and structure, i.e. $f(y; x, \theta_\alpha) \equiv f(y; x, \theta_\alpha, \mathcal{S}_\alpha)$. The meta-parameters can be learned through the maximum likelihood estimation(MLE) similar to [7]:

$$\hat{\psi} = \arg\max_{\psi} \prod_{\alpha=1,\ldots,L} \int_{\theta_\alpha} P(D_\alpha|\theta_\alpha, \mathcal{S}_\alpha) P(\theta_\alpha|\psi, \mathcal{S}_\alpha) d\theta_\alpha \tag{1}$$

and the parameters of a novel task can be estimated as

$$\hat{\theta}_* = \arg\max_{\theta_*} \prod_{(x^i, y^i) \in D_*} f(y^i; x^i, \theta_*, S_*) P(\theta|\psi, S_*) \tag{2}$$

where each task's model parameters depend on the task-specific parameter space $\mathbb{R}^{M_\alpha}$ as shown in Fig. 2.

B. Structures in Visual-tactile estimation

We can naturally instantiate object-specific structures for visual-tactile estimation problems. Before interacting with the object, we construct a structure $\mathcal{S}_\alpha$ consisting of a *computational mesh* $\mathcal{M}_\alpha$, *visual features* for elements in the mesh v_α. The computational mesh $\mathcal{M}_\alpha$ has n_α particles and m_α elements (subsets of particles). The ith particle has position p_i^t at time t.

At time step t, the robot executes *action* a^t (e.g., joint position commands) and it receives tactile and visual *observations* z^t (e.g. F/T readings and RGBD images). In the meta-learning framework, a touch comprises an input-output pair $x = (a^1, \ldots, a^T)$ and $y = (z^1, \ldots, z^T)$. The object state s^t is particle displacements $\{p_i^t\}_{i=1,\ldots,n_\alpha}$.

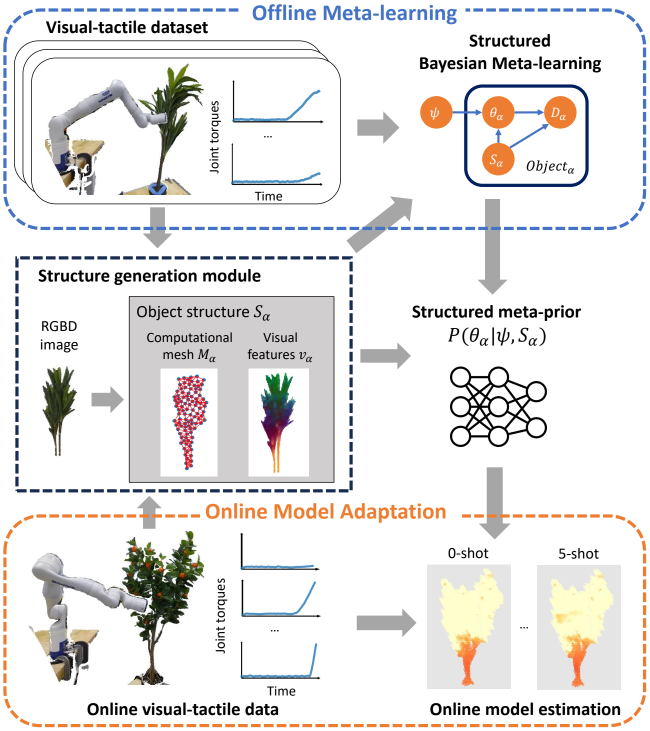

Fig. 2. An overview of the structured Bayesian meta-learning method.

In our formulation, θ_α denotes the *material parameters*, representing the object's time-invariant properties, such as Young's modulus. For simplicity, our implementation assumes that the state trajectory is a deterministic function of actions and material parameters, governed by the *dynamics model* $s^t = Dyn(s^{t-1}, a^t; \theta_\alpha)$ with known s^0. Finally, the *observation model* is given by $z^t \sim P(Z^t|s^t, \theta_\alpha, \mathcal{S}_\alpha)$. The touch observation likelihood $f(y; x, \theta, \mathcal{S})$ is simply a product of observation likelihoods:

$$f(y; x, \theta_\alpha, \mathcal{S}_\alpha) = \prod_{t=1}^{T} P(z^t|s^t, \theta_\alpha, \mathcal{S}_\alpha) \tag{3}$$

Here, the trajectory $s^0, \ldots, s^T$ is determined by rolling out the prediction $s^t = Dyn(s^{t-1}, a^t, \theta)$. Hence, the remaining part of implementing (1) and (2) is defining the prior $P(\theta_\alpha|\psi, \mathcal{S}_\alpha)$ that maps fixed-dimensional meta-parameters ψ to material parameters θ_α.

C. Efficient implementation with Gaussian likelihoods

We efficiently implement (1) and (2) for the Volumetric Stiffness Field (VSF) model and linear-Gaussian approximations of the observation model.

A VSF defines a volume of independent particles that resist displacements from their rest position with Hookean springs. This model has $\sim 10^5$ highly redundant parameters for densely sampled points. The structure defined by a VSF consists of a computational mesh $\mathcal{M}_\alpha$ with n_α particles and $m_\alpha = n_\alpha$ springs. Each particle responds to displacement with a Hookean reactive force $f_i^t = -K_{\alpha,i} \cdot (p_i^t - p_i^0)$, with

$K_{\alpha,i}$ the stiffness of this spring. The material parameters are the stiffness of each spring $\theta_\alpha = \{K_{\alpha,i}\}_{i=1,\dots,n}$. The VSF implementation leads to a *material-independent* simulation, i.e. $s_\alpha^t = Dyn(s_\alpha^{t-1}, a_\alpha^t, \mathcal{S}_\alpha)$, where $s_t \in \mathbb{R}^{3n_\alpha}$ represents the particle displacements.

We consider tactile observations z^t to be joint torques τ^t or 1-d pressure readings, which are linear observations of force f_i^t as well as $K_{\alpha,i}$. Assuming Gaussian noise in z^t, the online MAP problem (2) becomes a quadratic program (QP) solved using CVXPY [3]. We consider visual features mapped to each particle $v_{\alpha,i}$ and use Gaussian prior $K_{\alpha,i} \sim \mathcal{N}(\mu_\psi(v_{\alpha,i}), \sigma_\psi^2(v_{\alpha,i}))$. Evaluating (1) requires an integration over material parameters θ_α, which has closed form because the transition and observation models are linear-Gaussian using standard methods from Kalman filtering. The optimization problem was implemented using differentiable operations in PyTorch [12] and the Adam [9] optimizer.

III. Experiments and Results

A. Benchmark datasets

We acquire two benchmark datasets: 1) *plants* with 12 artificial plants and 7-d joint torques reading on Kinova Gen3 robot arm, and 2) *shoes* with 23 shoes and 1-d Punyo sensor pressure reading [1]. Several thousand touches are generated uniformly at random and each dataset is split into training, validation, and testing sets with multiple objects. We experiment with different training datasets, *broad* with all training examples and *narrow* with objects similar to test objects as in Fig. 3.

B. Implementations

VSF particles are sampled from the object's surface and interior [18]. We use pre-trained DINOv2 [11] to generate dense visual features v_α, and the image space visual features are projected to 3D space and linearly interpolated. The meta-prior is a multi-layer perception (MLP) and outputs the mean and covariance of a Gaussian distribution. The meta-parameters ψ are weights of qithe neural network. The joint torques τ^t are computed using Jacobian transpose, and pressure δp^t is proportional to the sum of each point's force magnitude.

We compare our method with two baselines: a) *Non-structured meta-learning* with tactile data only and no object-specific structure. An MLP takes arm joint angles a^t and outputs a tactile observation τ^t or δp^t, meta-trained and adapted using iMAML [13]; b) *Structured model with naïve prior* that uses "vanilla" VSF model without vision where prior assigns the same Gaussian distribution to all particles. The mean and variance are from the offline meta-learning dataset.

C. Qualitative and Quantitative Results

We train SBML on the plants and shoes training sets, and qualitatively visualize the zero-shot prediction and adaptation in Fig. 1 and Fig. 3. We observe that the zero-shot stiffness estimates qualitatively correspond to our intuition, with stiffer estimates at the trunk of plants and the toes of shoes. SBML performs best when the training and testing distributions closely align, as shown in shoes' tongue stiffness prediction.

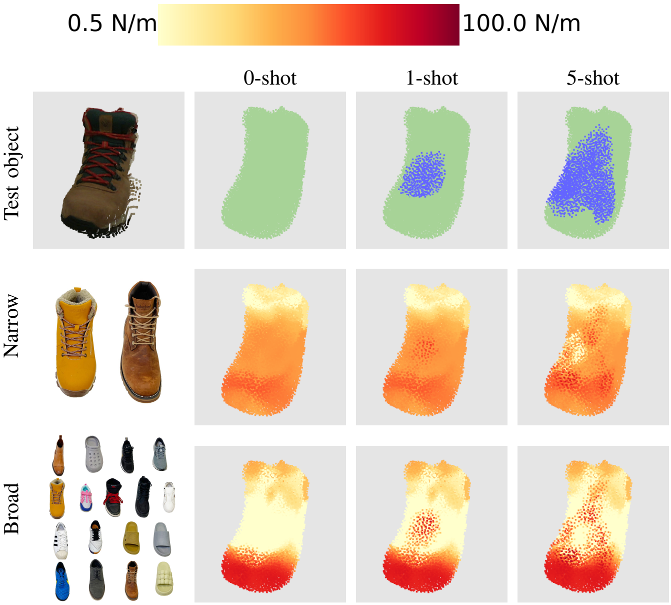

Fig. 3. Effects of the meta-training dataset on k-shot predictions. The first row after the test object shows the touched region (blue). The stiffness color map is in the log scale. We only show one object in the test dataset.

We show quantitative prediction accuracy on the Plants test dataset in Tab. I. The evaluation metric is prediction accuracy on a *query set* disjoint from the support set, in Nm for joint torques. The last row shows the VSF prediction error with many touches (>34 touches). The SBML prior has lower zero-shot prediction errors and quickly improves with online support data, achieving near-max-shot accuracy by 10 shots. This underscores the value of incorporating visual information to learn object material properties. In contrast, the iMAML baseline has slow adaptation without knowledge of object structure, the vanilla VSF has much worse zero- and few-shot prediction accuracy with naïve Gaussian prior. Both baselines have 10-shot prediction accuracy far from the many-shot result. Using different meta-training datasets for plants has no significant difference.

TABLE I
Joint torque prediction error (Nm) over query set on Plant benchmark test set objects.

	iMAML	Gaussian prior	SBML prior	SBML prior
	(broad dataset)	(average)	(narrow dataset)	(broad dataset)
0-shot	3.95 ± 0.85	5.23 ± 0.00	2.82 ± 0.04	2.80 ± 0.05
1-shot	5.71 ± 3.52	5.04 ± 0.04	2.78 ± 0.03	2.76 ± 0.05
5-shot	5.10 ± 2.16	4.42 ± 0.07	2.71 ± 0.05	2.67 ± 0.05
10-shot	4.40 ± 0.93	3.88 ± 0.09	2.66 ± 0.03	2.61 ± 0.04
VSF many-shot 2.57				

IV. Conclusion

We introduced a novel structured Bayesian meta-learning (SBML) approach for efficient visual-tactile model estimation, enhancing prediction accuracy on new objects by transferring offline knowledge from objects with various sizes and shapes. Our method surpasses non-informed and non-structured meta-learning techniques in zero- and few-shot accuracy, as shown by tests on plant and shoe datasets.

References

[1] Alex Alspach, Kunimatsu Hashimoto, Naveen Kuppuswamy, and Russ Tedrake. Soft-bubble: A highly compliant dense geometry tactile sensor for robot manipulation. In *2019 2nd IEEE International Conference on Soft Robotics (RoboSoft)*, pages 597–604, 2019. doi: 10.1109/ROBOSOFT.2019.8722713.

[2] Mark J Van der Merwe, Youngsun Wi, Dmitry Berenson, and Nima Fazeli. Integrated Object Deformation and Contact Patch Estimation from Visuo-Tactile Feedback. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.080.

[3] Steven Diamond, Eric Chu, and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization, version 0.2. http://cvxpy.org/, May 2014.

[4] Zackory Erickson, Alexander Clegg, Wenhao Yu, Greg Turk, C. Karen Liu, and Charles C. Kemp. What does the person feel? learning to infer applied forces during robot-assisted dressing. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6058–6065, 2017. doi: 10.1109/ICRA.2017.7989718.

[5] Zackory Erickson, Henry M Clever, Greg Turk, C Karen Liu, and Charles C Kemp. Deep haptic model predictive control for robot-assisted dressing. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4437–4444. IEEE, 2018.

[6] Barbara Frank, Cyrill Stachniss, Rüdiger Schmedding, Matthias Teschner, and Wolfram Burgard. Learning object deformation models for robot motion planning. *Robotics and Autonomous Systems*, 62(8):1153–1174, 2014. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2014.04.005.

[7] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ_UL-k0b.

[8] Advait Jain, Marc D Killpack, Aaron Edsinger, and Charles C Kemp. Reaching in clutter with whole-arm tactile sensing. *The International Journal of Robotics Research*, 32(4):458–482, 2013. doi: 10.1177/0278364912471865.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

[10] Mark Van der Merwe, Dmitry Berenson, and Nima Fazeli. Learning the dynamics of compliant tool-environment interaction for visuo-tactile contact servoing. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 2052–2061. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/merwe23a.html.

[11] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch]: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Conference on Neural Information Processing Systems*, pages 8024–8035, 2019.

[13] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/072b030ba126b2f4b2374f342be9ed44-Paper.pdf.

[14] Haochen Shi, Huazhe Xu, Zhiao Huang, Yunzhu Li, and Jiajun Wu. RoboCraft: Learning to See, Simulate, and Shape Elasto-Plastic Objects with Graph Networks. In *Proceedings of Robotics: Science and Systems*, New York City, NY, USA, June 2022. doi: 10.15607/RSS.2022.XVIII.008.

[15] Haochen Shi, Huazhe Xu, Samuel Clarke, Yunzhu Li, and Jiajun Wu. Robocook: Long-horizon elasto-plastic object manipulation with diverse tools. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=69y5fzvaAT.

[16] Yufei Wang, David Held, and Zackory Erickson. Visual haptic reasoning: Estimating contact forces by observing deformable object interactions. *IEEE Robotics and Automation Letters*, 7(4):11426–11433, 2022. doi: 10.1109/LRA.2022.3199684.

[17] Youngsun Wi, Andy Zeng, Peter R. Florence, and Nima Fazeli. Virdo++: Real-world, visuo-tactile dynamics and perception of deformable objects. In *Conference on Robot Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:252762347.

[18] Shaoxiong Yao and Kris Hauser. Estimating tactile models of heterogeneous deformable objects in real time. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12583–12589, 2023. doi: 10.1109/ICRA48891.2023.10160731.

[19] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e1021d43911ca2c1845910d84f40aeae-Paper.pdf.