Physics-aware Spatiotemporal Modules with Auxiliary Tasks for Meta-Learning

Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, Yan Liu
Department of Computer Science
University of Southern California
{sungyons,chuizhem,sirishar,yanliu.cs}@usc.edu

Abstract

Modeling the dynamics of real-world physical systems is critical for spatiotemporal prediction tasks, but challenging when data is limited. Although the knowledge of governing partial differential equations (PDE) of the data can be helpful for the fast adaptation to few observations, it is mostly infeasible to exactly find the equation for observations in real-world physical systems. In this work, we propose a framework, physics-aware meta-learning with auxiliary tasks whose spatial modules incorporate PDE-independent knowledge and temporal modules utilize the generalized features from the spatial modules to be adapted to the limited data, respectively. The framework is inspired by a local conservation law expressed mathematically as a continuity equation and does not require the exact form of governing equation to model the spatiotemporal observations. We apply the proposed framework to the real-world spatiotemporal prediction tasks and demonstrate its superior performance with limited observations.

1 Introduction

Deep learning has recently shown promise to play a major role in devising new solutions to applications with natural phenomena, such as climate change [1, 2], ocean dynamics [3], air quality [4, 5, 6], and so on. Deep learning techniques inherently require a large amount of data for effective representation learning, so their performance is significantly degraded when there are only a limited number of observations. However, in many tasks in physical systems in the real-world we only have access to a limited amount of data. One example is air quality monitoring [7], in which the sensors are irregularly distributed over the space – many sensors are located in urban areas whereas there are much fewer sensors in vast rural areas. Another example is extreme weather modeling and forecasting, i.e., temporally short events (e.g., tropical cyclones [8]) without sufficient observations over time. Thus, achieving robust performance from a few spatiotemporal observations in physical systems remains an essential but challenging problem.

Learning on a limited amount of data from physical systems can be considered as a few shot learning. Recently many meta-learning techniques [9, 10, 11, 12, 13, 14] have been developed to address this few shot learning setting, however, there are still some challenges for the existing meta-learning methods to be applied in modeling natural phenomena. First, it is not easy to find a set of similar meta-tasks which provide shareable latent representations needed to understand targeted observations. For instance, while image-related tasks (object detection [15] or visual-question-answering tasks [16, 17]) can take advantage of an image-feature extractor pre-trained by a large set of images [18] and well-designed architecture [19, 20, 21], there is no such large data corpus that is widely applicable for understanding natural phenomena. Second, exact equations behind natural phenomena are usually unknown, leading to the difficulty in reproducing the similar datasets via simulation.

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

In this work, we propose physics-aware modules designed for meta-learning to tackle the few shot learning challenges in physical observations. One of fundamental equations in physics describing the transport of physical quantity over space and time is a continuity equation:

$$\dot{\rho} + \nabla \cdot J = \sigma, \tag{1}$$

where ρ is the amount of the target quantity (u) per unit volume, J is the flux of the quantity, and σ is a source or sink, respectively. This fundamental equation can be used to derive more specific transport equations such as the convection-diffusion equation and Navier-Stokes equations. Thus, the continuity equation is the starting point to model spatiotemporal (conservative) observations which are accessible from sensors. Based on the form of ρ and J with respect to a particular quantity u, Eq. 1 can be generalized as:

$$\dot{u} = F(\nabla u, \nabla^2 u, \dots), \tag{2}$$

where the function $F(\cdot)$ describes how the target u is changed over time from its spatial derivatives. Inspired by the form of Eq. 2, we propose two modules: spatial derivative modules (SDM) and time derivative modules (TDM). Since the spatial derivatives such as ∇, $\nabla \cdot$, and ∇^2 are commonly used across different PDEs, the spatial modules are PDE-independent and they can be meta-initialized from synthetic data. Then, the PDE-specific temporal module is trained to learn the unknown function $F(\cdot)$ from few observations in the real-world physical systems. Based on the modularized PDEs, we introduce a novel approach that marries physics knowledge in spatiotemporal prediction tasks with meta-learning by providing shareable modules across spatiotemporal observations in the real-world.

Our contributions are summarized below:

- **Modularized PDEs and auxiliary tasks:** Inspired by forms of PDEs in physics, we decompose PDEs into shareable (spatial) and adaptation (temporal) parts. The shareable one is PDE-independent and specified by auxiliary tasks: supervision of spatial derivatives.
- **Physics-aware meta-learning:** We provide a framework for physcis-aware meta-learning, which consists of PDE-independent/-specific modules. The framework is flexible to be applied to the modeling of different or unknown dynamics.
- **Synthetic data for shareable modules:** We extract shareable parameters in the spatial modules from synthetic data, which can be generated from different dynamics easily.

Related work Since physics-informed neural networks are introduced in [22], which find that a solution of a PDE can be discovered by neural networks, physical knowledge has been used as an inductive bias for deep neural networks. Advection-diffusion equation is incorporated with deep neural networks for sea-surface temperature dynamics [23]. [24, 25] show that Lagrangian/Hamiltonian mechanics can be imposed to learn the equations of motion of a mechanical system and [26] regularizes a graph neural network with a specific physics equation. Rather than using explicitly given equations, physics-inspired inductive bias is also used for reasoning dynamics of discrete objects [27, 28] and continuous quantities [29]. [30, 31] propose a numeric-symbolic hybrid deep neural network designed to discover PDEs from observed dynamic data. While there are many physics-involved works, to the best of our knowledge, we are the first to provide a framework to use the physics-inspired inductive bias under the meta-learning settings to tackle the limited data issue which is pretty common for real-world data such as extreme weather events [8].

2 Physics-aware Meta-Learning with Auxiliary Tasks

2.1 Spatial Derivative Module

As we focus on the modeling and prediction of sensor-based observations, where the available data points are inherently on a spatially sparse irregular grid, we use graph networks for each module ϕ_k to learn the finite difference coefficients [32]. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, \dots, N\}$ and $\mathcal{E} = \{(i, j) : i, j \in \mathcal{V}\}$, a node i denotes a physical location $\mathbf{x}_i = (x_i, y_i)$ where a function value $u_i = u(x_i, y_i)$ is observed. Then, the graph signals with positional relative displacement as edge features are fed into the spatial modules to approximate spatial derivatives at every nodes.

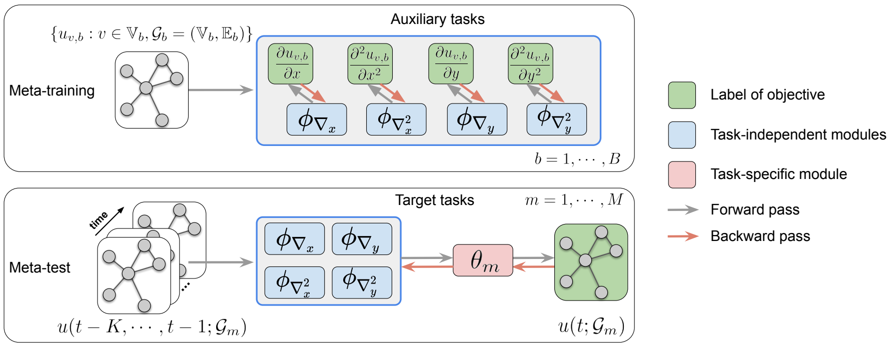


Figure 1: Schematic overview of the physics-aware meta-learning (PiMetaL).

2.2 Time Derivative Module

Once spatial derivatives are approximated, another learnable module is required to combine them for a target task. We use a recurrent graph network [33] for TDM.

2.3 Meta-Learning with Auxiliary Objective

In this section, we propose a physics-aware meta-learning framework to meta-initialize a spatial module by leveraging synthetic dataset with auxiliary tasks to provide reusuable features for the main tasks. The meta-initialization with the auxiliary tasks from synthetic datasets is particularly important. First, the spatial modules can be universal feature extractors for modeling observations following unknown physics-based PDEs. We propose that the PDE-independent spatial modules can be applicable as feature extractors across different dynamics as long as the dynamics follow a local form of conservation laws. Second, we can utilize synthetic data to meta-train the spatial modules as they are PDE-agnostic. This property allows us to utilize a large amount of synthetic datasets which are readily generated by numerical methods regardless of the exact form of PDE for targeted observations. Finally, we can provide a stronger inductive bias which is beneficial for modeling real-world observations but not available in the observations explicitly.

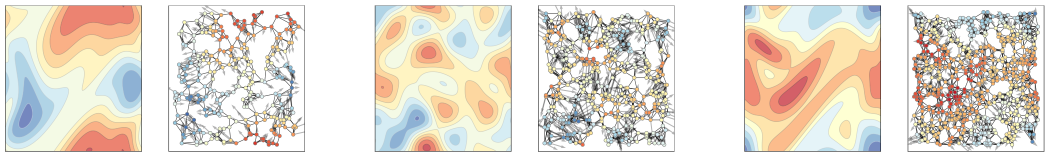
Algorithm 1 Meta-initialization with auxiliary tasks: Supervision of spatial derivatives

Input: A set of meta-train task datasets $D = fD_1, \ldots, D_B g$ where $D_b = (D_b^{tr}, D_b^{te})$. $D_b = f(u_i^b, e_{ij}^b, y_i^{(a_1,b)}, \ldots, y_i^{(a_K,b)}) : i \ 2 \ V_b, (i,j) \ 2 \ E_b g$ where $y_i^{(a_k,\,)}$ is an k-th auxiliary task label for the i-th node, given node/edge feature u^b and e^b, respectively. Learning rate α and β.
Output: Meta-initialized spatial modules .

1: Initialize spatial derivative modules $= (\phi_1, \ldots, \phi_K)$
2: **while** not converged **do**
3: **for** D_b in D **do**
4: ${}^{0}_{b} = \quad \alpha\, r \quad \sum_{k=1}^{K} L_k^{aux}(D_b^{tr}; \phi_k)$
5: **end for**
6: $\beta\, r \quad \sum_{b=1}^{B} \sum_{k=1}^{K} L_k^{aux}(D_b^{te}; \phi_{b,k}^{0})$
7: **end while**

Alg. 1 describes how the spatial modules are meta-initialized by MAML under the supervision of K different spatial derivatives. First, we generate values and spatial derivatives on a 2D regular grid from an analytical function. Then, we sample a finite number of points from the regular grid to represent discretized nodes and build a graph from the sampled nodes. Each graph signal and its discretization becomes input feature of a meta-train task and corresponding spatial derivatives are the auxiliary task labels. Fig. 2 visualizes graph signals and spatial derivatives for meta-initialization.

Once the spatial modules ( ) are initialized throughout meta-training, we reuse the modules for meta-test where the temporal module (the head of the network) are adapted on few observations from real-world sensors. We only adapt the head layer (θ) as like almost-no-inner-loop method



(a) (N, E, F)=(125,4,1) (b) (N, E, F)=(450,6,5) (c) (N, E, F)=(800,8,3)

Figure 2: Examples of generated spatial function values and graph signals. Node and edge features (function value and relative displacement, respectively) are used to approximate spatial derivatives (arrows). We can adjust the number of nodes (N, spatial resolution), the average number of edges (E, discretization), and the degree of fluctuation (F, scale of derivatives) to differentiate meta-train tasks.

in [34]: $\theta^0_m = \theta_m \quad \alpha \, r_{\theta_m} L(D^{tr}_m; \quad , \theta_m)$ and evaluate the model on D^{te}_m where $D_m = (D^{tr}_m, D^{te}_m)$ is a meta-test dataset.

3 Experimental Evaluation

Task: We adopt a set of multi-step spatiotemporal sequence generation tasks to evaluate our proposed framework. In each task, the data is a sequence of L frames, where each frame is a set of observations on N nodes in space. Then, we train an auto-regressive model with the first T frames (T-shot) and generate the following $L \quad T$ frames repeatedly from a given initial input (T-th frame) to evaluate its performance.

Datasets: For all experiments, we generate meta-train tasks with different spatial resolution, discretization, and fluctuation (Fig. 2) and the target observations are 2 real-world datasets: (1) **AQI-CO**: national air quality index (AQI) observations [7]; (2) **ExtremeWeather**: the extreme weather dataset [8]. For the AQI-CO dataset, we construct 12 meta-test tasks with the carbon monoxide (CO) ppm records from the first week of each month in 2015 at land-based stations. For the extreme weather dataset, we select the top-10 extreme weather events with the longest lasting time from the year 1984 and construct a meta-test task from each event with the observed surface temperatures at randomly sampled locations. Since each event lasts fewer than 20 frames, each task has a very limited amount of available data. In both datasets, graph signals are univariate.

Table 1: Multi-step prediction results (MSE) and standard deviations on the two real-world datasets.

T-shot	Method	AQI-CO		ExtremeWeather	
5-shot	FDM+RGN (scratch)	0.0291	0.0039	0.9883	0.5567
	PA-DGN (scratch)	0.0363	0.0090	0.9653	0.1384
	PiMetaL (meta-init)	**0.0253**	**0.0055**	**0.9167**	**0.0746**
7-shot	FDM+RGN (scratch)	0.0258	0.0023	0.7626	0.0602
	PA-DGN (scratch)	0.0225	0.0018	0.7478	0.0199
	PiMetaL (meta-init)	**0.0182**	**0.0019**	**0.7274**	**0.0089**
10-shot	FDM+RGN (scratch)	0.0213	0.0013	0.7090	0.0030
	PA-DGN (scratch)	0.0146	0.0005	0.4156	0.0145
	PiMetaL (meta-init)	**0.0115**	**0.0004**	**0.4066**	**0.0247**

Baselines: We evaluate the performance of a physics-aware architecture (**PA-DGN**) [29], which also consists of spatial derivative modules and recurrent graph networks (RGN), to see how the additional spatial information affects prediction performance for same architecture. Note that PA-DGN has same modules in PiMetaL and the difference is that PiMetaL utilizes meta-initialized spatial modules and PA-DGN is randomly initialized for learning from scratch on meta-test tasks. Additionally, the spatial modules in PA-DGN is replaced by finite difference method (**FDM+RGN**) to see if the numerical method provides better PDE-agnostic representations. The baselines and PiMetaL are trained on the meta-test support set only to demonstrate how the additional spatial information is beneficial for few-shot learning tasks.

Discussion: Table 1 shows the multi-step prediction performance of our proposed framework against the baselines on real-world datasets. Overall, PA-DGN and PiMetaL show similar trend such that the prediction error is decreased as longer series are available for few-shot adaptation. There are two important findings: first, with the similar expressive power in terms of the number of learnable

Table 2: Graph signal regression results (MSE, 10^{-3}) and standard deviations on the two regions of weather stations.

T-shot (Region)	GCN		GAT		GraphSAGE		GN		PA-DGN		PiMetaL	
5-shot (USA)	2.742	0.120	2.549	0.115	2.128	0.146	2.252	0.131	1.950	0.152	**1.794**	**0.130**
10-shot (USA)	2.371	0.095	2.178	0.066	1.848	0.206	1.949	0.115	1.687	0.104	**1.567**	**0.103**
5-shot (EU)	1.218	0.218	1.161	0.234	1.165	0.248	1.181	0.210	0.914	0.167	**0.781**	**0.019**
10-shot (EU)	1.186	0.076	1.142	0.070	1.044	0.210	1.116	0.147	0.831	0.058	**0.773**	**0.014**

parameters, the meta-initialized spatial modules provide high quality representations which are easily adaptable across different spatiotemporal dynamics in the real-world. This performance gap demonstrates that we can get a stronger inductive bias from synthetic datasets without knowing PDE-specific information. Second, the contribution of the meta-initialization is more significant when the length of available sequence is shorter ($T = 5$) and this demonstrates when the meta-initialization is particularly effective. Finally, the finite difference method provides proxies of exact spatial derivatives and the representations are useful particularly when $T = 5$ but its performance is rapidly saturated and it comes from the gap between the learnable spatial modules and fixed numerical coefficients. The results provide a new point of view on how to utilize synthetic or simulated datasets to handle challenges caused by limited datasets.

3.1 Graph Signal Regression

Task, datasets, and baselines: [35] conducted a graph signal regression task: predict the temperature x_t from the temperature on the previous 5 days ($x_{t-5} : x_{t-1}$). We split the **GHCN** dataset[1] spatially into two regions: (1) the USA (1,705 stations) and (2) Europe (EU) (703 stations) where there are many weather stations full functioning. In this task, the number of shots is defined as the number of input and output pairs to train a model. As the input length is fixed, more variants of graph neural networks are considered as baselines. We concatenate the 5-step signals and feed it into Graph convolutional networks (**GCN**) [36], Graph attention networks (**GAT**) [37], **GraphSAGE** [38], and Graph networks (**GN**) [39] to predict next signals across all nodes.

Discussion: Table 2 shows the results of the graph signal regression task across different baselines and the proposed method. There are two patterns in the results. First, although in general we observe an improvement in performance for all methods when we move from the 5-shot setting to the 10-shot setting, PiMetaL's performance yields the smallest error. Second, for the EU dataset, while 5-shot seems enough to achieve stable performance, it demonstrates that the PDE-independent representations make the regression error converge to a lower level. Overall, the experimental results prove that the learned spatial representations from simulated dynamics are beneficial for learning on limited data.

4 Conclusion

In this paper, we propose a framework for physics-aware meta-learning with auxiliary tasks. By incorporating PDE-independent/-invariant knowledge (spatial derivatives) from simulated data, the framework provide reusable features to meta-test tasks with a limited amount of data. Experiments show that auxiliary tasks and physics-aware meta-learning help construct reusable modules that improve the performance of spatiotemporal predictions in real-world tasks where data is limited. Although introducing auxiliary tasks based on synthetic datasets improves the prediction performance, they need to be chosen and constructed manually and intuitively. Designing and identifying the most useful auxiliary tasks and data will be the focus of our future work.

[1] Global Historical Climatology Network (GHCN) provided by National Oceanic and Atmospheric Administration (NOAA). https://www.ncdc.noaa.gov/ghcn-daily-description

Broader Impact

The purpose of this work is to provide a new approach for robust and efficient learning from limited observations without knowing the governing equations or physics rules in real-world. It has been considered that natural phenomena are not appropriate to be used in meta-learning as different observations have different dynamics and it is hard to find shareable representations across various phenomena. Our work explores a way to address this challenge via modularized partial differential equations and meta-learning. Since it is common to see the limited observations of natural phenomena, our work provides a methodology about utilizing synthetic data to address few shot learning for spatiotemporal observations.

References

[1] A Manepalli, A Albert, A Rhoades, D Feldman, and M Prabhat. Emulating numeric hydroclimate models with physics-informed conditional generative adversarial networks. *Environmetrics*, 2019.

[2] Ján Drgona, Lieve Helsen, and Draguna Vrabie. Stripping off the implementation complexity of physics-based model predictive control for buildings via deep learning. *https://www. nips. cc/*, 2019.

[3] Gautier Cosne, Guillaume Maze, and Pierre Tandeo. Coupling oceanic observation systems to study mesoscale ocean dynamics. *arXiv preprint arXiv:1910.08573*, 2019.

[4] Ping-Wei Soh, Jia-Wei Chang, and Jen-Wei Huang. Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. *Ieee Access*, 6:38186–38199, 2018.

[5] Shengdong Du, Tianrui Li, Yan Yang, and Shi-Jinn Horng. Deep air quality forecasting using hybrid deep learning framework. *arXiv preprint arXiv:1812.04783*, 2018.

[6] Yijun Lin, Nikhit Mago, Yu Gao, Yaguang Li, Yao-Yi Chiang, Cyrus Shahabi, and José Luis Ambite. Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 359–368, 2018.

[7] Lex Berman. National aqi observations (2014-05 to 2016-12). Harvard Dataverse, 2017.

[8] Evan Racah, Christopher Beckham, Tegan Maharaj, Samira Ebrahimi Kahou, Mr Prabhat, and Chris Pal. Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. In *Advances in Neural Information Processing Systems*, pages 3402–3413, 2017.

[9] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: The meta-meta-... hook*. Diplomarbeit, Technische Universität München, München, 1987.

[10] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in neural information processing systems*, pages 3981–3989, 2016.

[11] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017.

[12] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1842–1850, 2016.

[13] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[16] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[17] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[22] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[23] Emmanuel de Bezenac, Arthur Pajot, and Patrick Gallinari. Deep learning for physical processes: Incorporating prior scientific knowledge. In *International Conference on Learning Representations*, 2018.

[24] Michael Lutter, Christian Ritter, and Jan Peters. Deep lagrangian networks: Using physics as model prior for deep learning. *arXiv preprint arXiv:1907.04490*, 2019.

[25] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, pages 15353–15363, 2019.

[26] Sungyong Seo and Yan Liu. Differentiable physics-informed graph networks. *arXiv preprint arXiv:1902.02950*, 2019.

[27] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pages 4502–4510, 2016.

[28] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016.

[29] Sungyong Seo, Chuizheng Meng, and Yan Liu. Physics-aware difference graph networks for sparsely-observed dynamics. In *International Conference on Learning Representations*, 2020.

[30] Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. Pde-net: Learning pdes from data. *arXiv preprint arXiv:1710.09668*, 2017.

[31] Zichao Long, Yiping Lu, and Bin Dong. Pde-net 2.0: Learning pdes from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, 399:108925, 2019.

[32] Yohai Bar-Sinai, Stephan Hoyer, Jason Hickey, and Michael P Brenner. Learning data-driven discretizations for partial differential equations. *Proceedings of the National Academy of Sciences*, 116(31):15344–15349, 2019.

[33] Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin Riedmiller, Raia Hadsell, and Peter Battaglia. Graph networks as learnable physics engines for inference and control. *arXiv preprint arXiv:1806.01242*, 2018.

[34] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

[35] Michaël Defferrard, Martino Milani, Frédérick Gusset, and Nathanaël Perraudin. Deepsphere: a graph-based spherical cnn. In *International Conference on Learning Representations*, 2019.

[36] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[37] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[38] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.

[39] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.