Few-shot Malware Recognition

Ching-Yuan Peng, Jayant Dabas, Jinu Hyun, Wei-Cheng Hsu
Center for Data Science, New York University
{cp4516,jd5831,jh10305,wh2757}@nyu.edu

Abstract

Android malware evolves rapidly, producing new families long before labeled data becomes available and rendering supervised classifiers ineffective against zero-day threats. In this study, we address the "small data" problem in cybersecurity by exploring Few-Shot Learning (FSL) and Meta-Learning techniques for malware recognition. Unlike natural-image Zero-Shot Learning (ZSL), where semantics align closely with visual structure, malware visualization encodes behavioral and structural signals such as entropy, resulting in weak semantic alignment and extreme distribution shift. We propose a unified benchmark protocol and evaluation pipeline built on top of the MalVis dataset to compare optimization-based meta-learning methods such as MAML, MAML++, Reptile, and the enhanced Mi-MAML against classical ZSL approaches including SAE, PSR-ZSL, and DAZLE. Across few-shot, generalized zero-shot, and benign-dominant settings, Mi-MAML consistently delivers the strongest adaptation. Our findings highlight malware few-shot learning as a challenging departure from traditional spatial ZSL and establish MalVis-based episodic evaluation as a rigorous test for studying rapid generalization under severe obfuscation and abrupt class shifts.

Keywords: Few-shot learning, malware detection, meta-learning, zero-shot learning, MalVis

Introduction

The proliferation of Android malware presents a persistent threat to cybersecurity infrastructure. Modern malware developers increasingly employ sophisticated obfuscation techniques such as polymorphism and packing to alter the binary structure of malicious code without changing its underlying behavior. This rapid mutation renders traditional signature-based detection ineffective and poses a significant challenge for deep learning systems. While supervised learning models have achieved high accuracy in malware classification, they rely heavily on large-scale, annotated datasets. In real-world scenarios, when a new malware family emerges, gathering and labeling sufficient samples is a slow, expensive, and unreliable process. This creates a critical "Small Data" challenge: defense systems must learn to identify new threats from only a handful of examples.

To address this, our research is motivated by the need to detect unseen malware families fast and reliably. We adopt a visual approach to malware analysis, building upon the foundational work of Nataraj et al. (2011), who demonstrated that malware binaries can be mapped to grayscale images where texture patterns correspond to code structures. By treating malware detection as a computer vision problem, we can leverage visual similarities that remain robust despite minor code obfuscations. However, standard image classification is insufficient when the target classes (new families) are unknown during training.

Therefore, the primary goal of this research is to explore whether Meta-Learning and Few-Shot Learning (FSL) paradigms can effectively recognize entirely new malware categories with minimal data support. We conduct a comparative study on the MalVis dataset, evaluating the performance of diverse algorithms. Specifically, we benchmark optimization-based meta-learning methods, such as Model-Agnostic Meta-Learning (MAML) and Reptile, against semantic-based approaches like Domain-Aware Zero-Shot Learning (DAZLE). Our study aims to determine which learning strategy yields superior generalization capabilities in the high-stakes, low-data environment of zero-day malware detection.

Related Works

Malware Visualization

Traditional malware analysis relies on static disassembly or dynamic execution, which are computationally expensive and vulnerable to obfuscation. To mitigate this, Nataraj et al. (2011) proposed a novel approach by visualizing malware binaries as grayscale images. In their method, a binary file is treated as a vector of 8-bit unsigned integers and reshaped into a 2D matrix, where each byte represents a pixel intensity (0–255). This transformation maps code structures (e.g., .text, .data sections) into distinct visual textures. Consequently, standard computer vision techniques, such as GIST descriptors and Convolutional Neural Networks (CNNs), can be applied to classify malware families based on visual similarity rather than specific instruction sequences.

Few-Shot and Meta-Learning

Meta-learning, or "learning to learn," aims to develop models that can adapt to new tasks with minimal data. Optimization-based approaches have gained significant traction in this domain. Finn et al. (2017) introduced Model-Agnostic Meta-Learning (MAML), which learns a set of initialization parameters that are sensitive to task-specific changes. Antoniou et al. (2019) later proposed MAML++, stabilizing the training process by decoupling the learning rates. Nichol et al. (2018)

introduced Reptile, a first-order approximation that simplifies computation by moving initialization weights towards trained parameters.

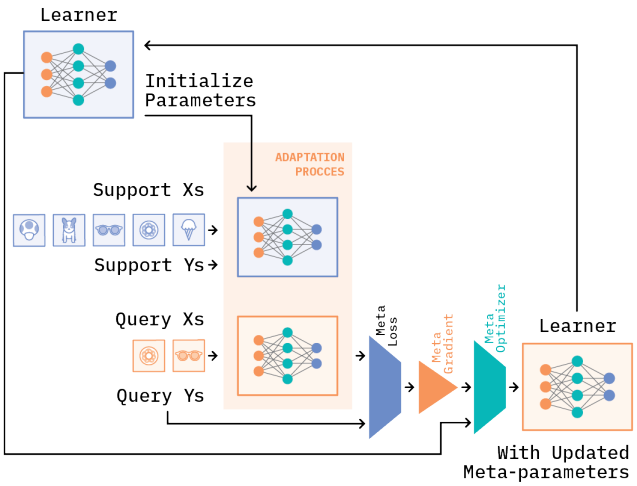

Figure 1: Meta Leraning Architecture

More recently, Ji et al. (2024) proposed **Mi-MAML (Multi-improved MAML)**, specifically designed for malware classification. Mi-MAML enhances the standard MAML framework by integrating data augmentation techniques (such as Lab color space transformation) and customizing the neural network architecture with adaptive learning rate schedules. This method directly addresses the overfitting issues common in high-dimensional malware imagery, making it a critical baseline for our study.

Zero-Shot Learning (ZSL)

Zero-Shot Learning (ZSL) extends recognition to classes not seen during training by leveraging semantic transfer. Methods often utilize semantic attributes or generative models to bridge the gap between visual features and class labels. For instance, the Semantic Autoencoder (SAE) (Kodirov et al., 2017) projects visual features into a semantic space, while Dense Attribute-Based Zero-Shot Learning (DAZLE) (Huynh & Elhamifar, 2020) utilizes attention mechanisms to align dense visual features with semantic attributes. However, these methods are predominantly evaluated on natural image datasets like CUB-200 (birds) or SUN (scenes), where classes have intuitive, "expertly annotated" attributes (e.g., wings, colors). In the malware domain, such semantic attributes are often abstract or non-existent, posing a unique challenge for applying conventional ZSL frameworks.

Methods

Dataset

We evaluate all models on MalVis, a large-scale Android malware visualization dataset released by Makkawy et al. (Makkawy et al., 2024) and derived from the AndroZoo corpus (Allix et al., 2016). We use the dataset's provided 256×256 RGB visual representations, in which the red and blue channels encode sliding-window entropy and the green channel encodes either class-byte semantics or N-gram structural patterns. Following prior findings indicating stronger generalization, we adopt the **MalVis-B** (N-gram) variant, which captures underlying structural patterns and contextual dependencies.

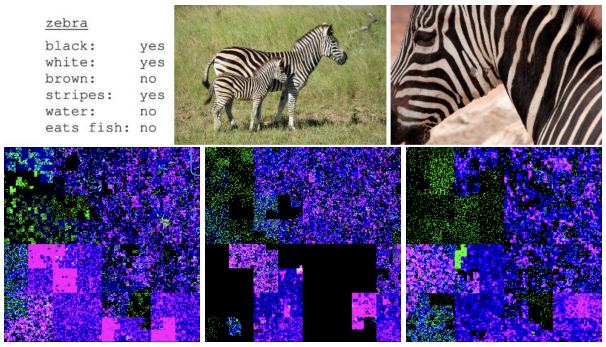

Figure 2: ImageNet vs MalVis-B malware Dataset

For evaluation, we follow standard practice by subsampling the dataset to represent small data and organizing malware families into disjoint training, validation, and testing partitions. All experiments operate on the original labels and representations provided by MalVis.

Framework

We frame malware detection as a few-shot image classification problem evaluated under controlled experimental settings. The overall pipeline consists of two stages: (1) using the visual representations provided by the MalVis dataset, and (2) applying an episodic evaluation strategy to assess model adaptation and generalization from limited labeled data.

Episodic Learning Formulation To simulate the "small data" scenario, we adopt the standard episodic learning framework commonly used in few-shot classification. We define a distribution over tasks $p(\mathcal{T})$, where each task $\mathcal{T}_i$ corresponds to classifying a subset of malware families.

Each task (or episode) consists of two disjoint sets:

- **Support Set ($\mathcal{S}$):** A small set of labeled examples used for model adaptation (e.g., 5 classes with 1 image each).
- **Query Set (Q):** A set of unlabeled samples from the same classes used to evaluate the adapted model.

Formally, in the N-way K-shot setting, the support set is defined as $\mathcal{S} = \{(x_k, y_k) \mid k = 1, 2, \ldots, (N \times K)\}$, where x_k is the malware image and y_k is the family label. The goal of the meta-learner is to minimize the prediction error on the query set Q after updating its parameters based on the information in $\mathcal{S}$. This forces the model to learn transferable features rather than memorizing specific class characteristics.

Models Evaluated

Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) is our foundational framework. MAML operates on a bilevel optimization scheme: an *inner loop* that adapts the model parameters θ to a specific task $\mathcal{T}_i$ via gradient descent, and an *outer loop* that updates the initial parameters to minimize the loss across a batch of tasks. Formally, the model seeks an initialization θ such that one or a few gradient steps on a new task will result in maximal performance.

Reptile serves as an efficient and simple meta-learning baseline. Its goal is to find a set of highly adaptable initial parameters (ϕ) that are centrally located among the optimal solutions for all tasks, thereby maximizing the model's ability to quickly adapt to new tasks. Reptile avoids the computational complexity of second-order derivatives by using an intuitive update rule. In the **Inner Loop**, the model trains on a specific task for a few steps, yielding adapted parameters ($\tilde{\phi}$). In the **Outer Loop**, the base parameters (ϕ) are simply moved towards these adapted parameters ($\tilde{\phi}$). This allows the initial parameters to quickly learn effective adaptation strategies across different tasks.

MAMl++ To address the training instability often observed in MAML, we also evaluate **MAML++** (Antoniou et al., 2019). This variant introduces several architectural improvements, including per-step learnable learning rates and query set multi-step loss optimization. These modifications stabilize the training trajectory and improve convergence speed, which is particularly beneficial for high-dimensional malware imagery.

Multi-Improved Model-Agnostic Meta-Learning (MI-MAML) is an enhanced variant of MAML specifically designed for Few-shot Classification. It addresses the training instability and slow convergence of MAML by integrating two main improvements: **Dynamic Inner-Loop Learning Rate (DILLR)**, which automatically adjusts the task adaptation learning rate α_t, and a **Momentum-based Meta-Update** for the outer loop. This momentum mechanism effectively utilizes historical gradient information to smooth the meta-optimization path, leading to improved convergence and stability over the standard MAML.

Semantic Autoencoder (SAE) (Kodirov et al., 2017) serves as a classical zero-shot learning baseline. SAE learns a linear encoder–decoder model that projects visual features into a semantic space while enforcing reconstruction consistency. Given visual features $X \in \mathbb{R}^{d\times n}$ and class semantic descriptors $S \in \mathbb{R}^{k\times n}$, the encoder maps X into the semantic space using a projection matrix W, and the decoder reconstructs back using tied weights $W^\top$.

The learning objective minimizes reconstruction error while constraining the semantic embedding:

$$\min_W \|X - W^\top S\|_F^2 \quad \text{s.t.} \quad S = WX.$$

Substituting the constraint yields the standard SAE formulation:

$$\min_W \|X - W^\top WX\|_F^2 + \lambda \|W\|_F^2,$$

where λ controls regularization.

At inference time, unseen samples are embedded via

$$\hat{s} = Wx,$$

and classified by nearest-neighbor matching in semantic space. SAE provides a strong semantic alignment baseline but is limited in malware imagery where pixel-level structure does not correspond to meaningful class attributes.

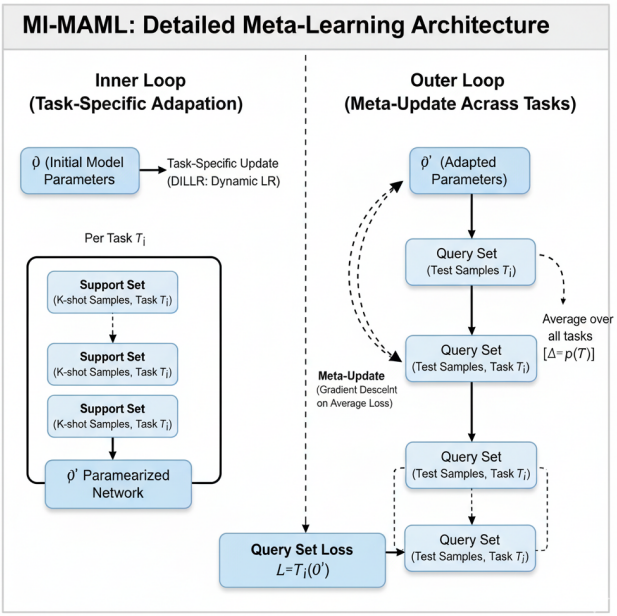

Figure 3: MI-MAML Architecture

Preserving Semantic Relations for Zero-Shot Learning (PSR-ZSL) (Biswas & Annadani, 2018) extends classical semantic embedding methods by explicitly modeling relational structure among classes. Instead of aligning visual features to individual semantic prototypes, PSR-ZSL preserves pairwise semantic distances to enforce meaningful neighborhood structure in the embedding space.

Let $A \in \mathbb{R}^{k\times C}$ denote attribute vectors for C classes and $X \in \mathbb{R}^{d\times n}$ the visual features. PSR-ZSL learns a mapping f_θ that projects attributes into the visual space:

$$\hat{X} = f_\theta(A),$$

and reconstructs attributes through a decoder g_ϕ:

$$\hat{A} = g_\phi(\hat{X}).$$

The objective combines attribute reconstruction with semantic relation preservation:

$$\min_{\theta,\phi} \|A - \hat{A}\|_F^2 + \gamma \sum_{i,j} \left(d(A_i, A_j) - d(\hat{X}_i, \hat{X}_j)\right)^2,$$

where $d(\cdot,\cdot)$ denotes Euclidean distance and γ controls the importance of relational consistency.

During inference, unseen class attributes are projected to the visual space via f_θ, and test features are assigned using nearest-neighbor matching. Although PSR-ZSL improves over naïve attribute-alignment approaches, it remains limited in malware imagery where semantic relationships between classes do not align with pixel-level structure.

DAZLE (Huynh & Elhamifar, 2020) is a discriminative attribute-based zero-shot learning model that leverages attribute activation maps to localize semantic concepts within convolutional feature maps. Given malware images lack human-interpretable semantics, DAZLE serves as a challenging upper-bound baseline in our study.

Given feature maps $F \in \mathbb{R}^{C\times H\times W}$ extracted from a backbone (e.g., ResNet-101), DAZLE learns attribute embeddings W_{att} and produces attribute activation scores:

$$a = W_{\text{att}}^{\top} \cdot \text{GAP}(F),$$

where $\text{GAP}(\cdot)$ denotes global average pooling. These attribute scores are combined to form a class compatibility score:

$$s_c = A_c^{\top} a,$$

where A_c is the attribute vector of class c.

Training optimizes a cross-entropy objective over seen classes, encouraging attributes to activate discriminatively:

$$\mathcal{L} = -\sum_i \log \frac{\exp(s_{y_i})}{\sum_c \exp(s_c)}.$$

At test time, unseen classes are scored using their attribute prototypes. While DAZLE performs well on natural image datasets with rich human-annotated attributes, its reliance on interpretable semantic descriptors makes it poorly suited for malware imagery, where semantics do not correspond to pixel-space attributes.

Experiments

Experimental Setup

N-way K-shot Protocol Following standard few-shot learning evaluation protocols, we adopt the N-way K-shot episodic setting. In our experiments, we fix $N = 5$ (5-way classification) to simulate a realistic scenario where an analyst must distinguish between a small set of potential malware families. To evaluate performance under varying degrees of data scarcity, we test two distinct settings:

- **1-shot:** The model is provided with only a single labeled example per class. This represents the most challenging "zero-day" detection scenario.
- **5-shot:** The model is provided with five labeled examples per class, allowing for slightly more statistical guidance.

Evaluation Scenarios To comprehensively assess the robustness of our models, we designed three distinct testing scenarios defined in Table 1. These settings allow us to evaluate the models' generalization capabilities across different data distributions and difficulty levels.

Table 1: Definition of the three experimental scenarios used to evaluate model robustness and generalization.

Test Scenario	Description & Objective
Only Unseen (100% Unseen Classes)	Measures pure Few-Shot Learning (FSL) performance: how well the model identifies completely novel malware families.
General (Partial) (15% Benign + All Unseen Classes)	Tests real-world deployment behavior: how well the model detects new malware while maintaining low false positives for benign traffic.
General (All) (All Seen Classes including Benign + All Unseen Classes)	Evaluates Generalized Few-Shot Learning (GFSL): the trade-off between recognizing known threats (seen) and novel threats (unseen).

Evaluation Metrics To ensure a fair assessment of the model's robustness, we adopt **Per-class Accuracy** as our primary evaluation metric. Unlike standard global accuracy, which can be skewed by class imbalances or dominant easy-to-classify samples, per-class accuracy calculates the correct classification rate for each specific malware family individually before averaging.

Reported results represent the mean accuracy computed over randomly sampled test episodes (e.g., 600 episodes). This metric is particularly critical in cybersecurity: a defense system must reliably detect *every* new variant, ensuring that no single evasive family is consistently misclassified due to statistical under-representation.

Baselines

We compare our approach against the following state-of-the-art methods. As mentioned in previous sections, we include both optimization-based and semantic-based approaches:

- **Optimization-based:** MAML (Finn et al., 2017), MAML++ (Antoniou et al., 2019), Reptile (Nichol et al., 2018), and Mi-MAML (Ji et al., 2024).
- **Semantic-based (ZSL):** SAE (Kodirov et al., 2017), PSR-ZSL (Biswas & Annadani, 2018), and DAZLE (Huynh & Elhamifar, 2020).

Experimental Results

Evaluation Protocol

We evaluate all models under both **1-shot** and **5-shot** settings to assess their adaptation capability under extreme and moderate data scarcity. All results are averaged over multiple episodes with fixed random seeds to ensure reproducibility.

Performance on Unseen Malware Families

Table 2 reports classification accuracy on the **Test Unseen** setting, where all malware families are unseen during meta-training. This evaluation measures pure few-shot transfer capability without interference from seen classes.

Table 2: Classification accuracy (%) on unseen malware families

Model	1-shot	5-shot
MAML	0.38	0.49
MAML++	0.42	0.48
Reptile	0.40	0.51
Mi-MAML	**0.49**	**0.55**
SAE	0.31	0.32
PSR-ZSL	0.42	0.45
DAZLE	0.33	0.44

Mi-MAML achieves the highest unseen accuracy under both 1-shot and 5-shot settings among all meta-learning approaches, indicating superior generalization to novel malware families. In contrast, semantic-based methods such as SAE and DAZLE perform notably worse, suggesting that semantic priors alone are insufficient for malware family generalization.

Generalized Few-Shot Malware Classification

We evaluate all models under the **generalized few-shot setting**, where both seen and unseen malware families are jointly considered at test time. Following standard practice in generalized few-shot and GZSL evaluation, we report the **Harmonic Mean (H)** of seen and unseen accuracy, which captures the trade-off between memorization and generalization.

Table 3: Generalized evaluation using Harmonic Mean (H)

Model	1-shot H	5-shot H
MAML	0.53	0.53
MAML++	0.55	0.64
Reptile	0.44	0.57
Mi-MAML	**0.61**	**0.68**
SAE	0.12	0.24
PSR-ZSL	0.60	0.64
DAZLE	0.42	0.49

Mi-MAML achieves the highest harmonic mean among meta-learning methods in both 1-shot and 5-shot settings, indicating a stronger balance between unseen adaptation and seen performance. While semantic-based methods can be competitive, their reliance on semantic priors limits robustness in malware domains with weak or noisy semantics. Overall, Mi-MAML demonstrates more stable and sample-efficient generalization in generalized few-shot malware detection.

Benign-Dominant Deployment Scenario

To better reflect real-world malware detection environments, we further evaluate all models under a **benign-dominant generalized setting**, where benign samples constitute the majority of seen data and unseen malware families appear sparsely at test time. This setting emphasizes robustness to distribution shift and the ability to detect novel malware without degrading performance on benign traffic.

Table 4: Benign-dominant generalized evaluation (1-shot)

Model	Seen (Benign)	Unseen
MAML	0.71	0.40
MAML++	0.78	0.47
Reptile	0.58	0.41
Mi-MAML	**0.85**	**0.52**
SAE	0.85	0.44
PSR-ZSL	0.76	0.46
DAZLE	0.60	0.34

Table 5: Benign-dominant generalized evaluation (5-shot)

Model	Seen (Benign)	Unseen
MAML	0.78	0.55
MAML++	0.82	0.58
Reptile	0.70	0.52
Mi-MAML	**0.93**	**0.62**
SAE	0.86	0.44
PSR-ZSL	0.84	0.55
DAZLE	0.59	0.44

Mi-MAML consistently achieves strong unseen malware detection while preserving high benign accuracy under both 1-shot and 5-shot settings, demonstrating superior robustness to class imbalance and distribution shift. In contrast, semantic-based methods remain competitive on benign samples but generalize less reliably to unseen malware, highlighting the advantage of task-adaptive meta-learning in benign-dominant deployments.

Summary of Results

Across all evaluation settings, Mi-MAML consistently outperforms standard meta-learning baselines on unseen and generalized malware detection tasks. The performance gains are particularly pronounced under extreme data scarcity (1-shot), highlighting Mi-MAML's suitability for real-world few-shot malware classification.

Discussion

The results highlight a key mismatch between malware visualization and traditional semantic-based zero-shot learning. Malware images do not contain clear, human-interpretable semantic attributes, which leads to weak alignment between visual patterns and semantic descriptors. In contrast, task-adaptive meta-learning is better suited to this setting because it allows models to quickly adapt to new malware families using very little labeled data. The strong performance of Mi-MAML suggests that stable meta-optimization is important for handling the variability and distribution shifts caused by malware obfuscation. Overall, the study demonstrates that task-adaptive meta-learning is a more appropriate paradigm than semantic-based zero-shot learning for malware recognition in low-data settings. Despite these gains, malware visualizations capture only partial behavioral information, which inherently limits performance even for adaptive models. Future work should explore representations that move beyond spatial image layouts toward structure-aware encodings, as well as incorporate attention mechanisms or information-guided feature extraction to better capture meaningful malware patterns under extreme data scarcity.

Acknowledgments

The authors acknowledge the usage of Claude, a large language model developed by Anthropic, in the preparation of this assignment. Claude was employed in the following manner within this assignment: grammatical correction.

References

Allix, K., Bissyandé, T. F., Klein, J., & Le Traon, Y. (2016). Androzoo: Collecting millions of android apps for the research community. *Proceedings of the 13th International Conference on Mining Software Repositories*, 468–471.

Antoniou, A., Edwards, H., & Storkey, A. (2019). How to train your maml. *International Conference on Learning Representations*.

Biswas, S., & Annadani, Y. (2018). Preserving semantic relations for zero-shot learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7603–7612.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *International conference on machine learning*, 1126–1135.

Huynh, D. T., & Elhamifar, E. (2020). Fine-grained generalized zero-shot learning via dense attribute-based attention. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4483–4493.

Ji, Y., Zou, K., & Zou, B. (2024). Mi-maml: Classifying few-shot advanced malware using multi-improved model-agnostic meta-learning. *Cybersecurity*, *7*(1), 1–14.

Kodirov, E., Xiang, T., & Gong, S. (2017). Semantic autoencoder for zero-shot learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4447–4456.

Makkawy, S. J., Alblwi, A. H., De Lucia, M. J., & Barner, K. E. (2024). Improving android malware detection with entropy bytecode-to-image encoding framework. *2024 33rd International Conference on Computer Communications and Networks (ICCCN)*, 1–9.

Nataraj, L., Karthikeyan, S., Jacob, G., & Manjunath, B. (2011). Malware images: Visualization and automatic classification. *Proceedings of the 8th International Symposium on Visualization for Cyber Security*, 1–7.

Nichol, A., Achiam, J., & Schulman, J. (2018). On first-order meta-learning algorithms.