

# Meta-Learning with Context-Agnostic Initialisations

Toby Perrett, Alessandro Masullo, Tilo Burghardt,  
Majid Mirmehdi, Dima Damen

Department of Computer Science, University of Bristol, UK  
[Toby.Perrett@bristol.ac.uk](mailto:Toby.Perrett@bristol.ac.uk)

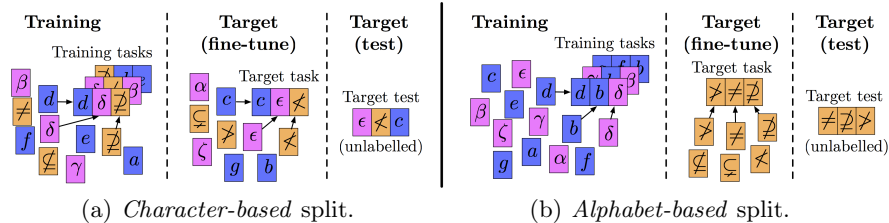
**Abstract.** Meta-learning approaches have addressed few-shot problems by finding initialisations suited for fine-tuning to target tasks. Often there are additional properties within training data (which we refer to as context), not relevant to the target task, which act as a distractor to meta-learning, particularly when the target task contains examples from a novel context not seen during training.

We address this oversight by incorporating a context-adversarial component into the meta-learning process. This produces an initialisation which is both context-agnostic and task-generalised. We evaluate our approach on three commonly used meta-learning algorithms and four case studies. We demonstrate our context-agnostic meta-learning improves results in each case. First, we report few-shot character classification on the Omniglot dataset, using alphabets as context. An average improvement of 4.3% is observed across methods and tasks when classifying characters from an unseen alphabet. Second, we perform few-shot classification on Mini-ImageNet, obtaining context from the label hierarchy, with an average improvement of 2.8%. Third, we perform few-shot classification on CUB, with annotation metadata as context, and demonstrate an average improvement of 1.9%. Fourth, we evaluate on a dataset for personalised energy expenditure predictions from video, using participant knowledge as context. We demonstrate that context-agnostic meta-learning decreases the average mean square error by 30%.

## 1 Introduction

Current deep neural networks require significant quantities of data to train for a new task. When only limited labelled data is available, meta-learning approaches train a network initialisation on other *source* tasks, so it is suitable for fine-tuning to new few-shot *target* tasks [1]. Often, training data samples have additional properties, which we collectively refer to as *context*, readily available through metadata. We give as an example the *alphabet* in a few-shot character recognition task (Fig. 1). This is distinct from multi-label problems as we pursue invariance to the context (i.e. alphabet), so as to generalise to unseen contexts in fine-tuning, rather than predicting its label.

In this work, we focus on problems where the target task is not only novel but does not have the same context as tasks seen during training. This is a difficult

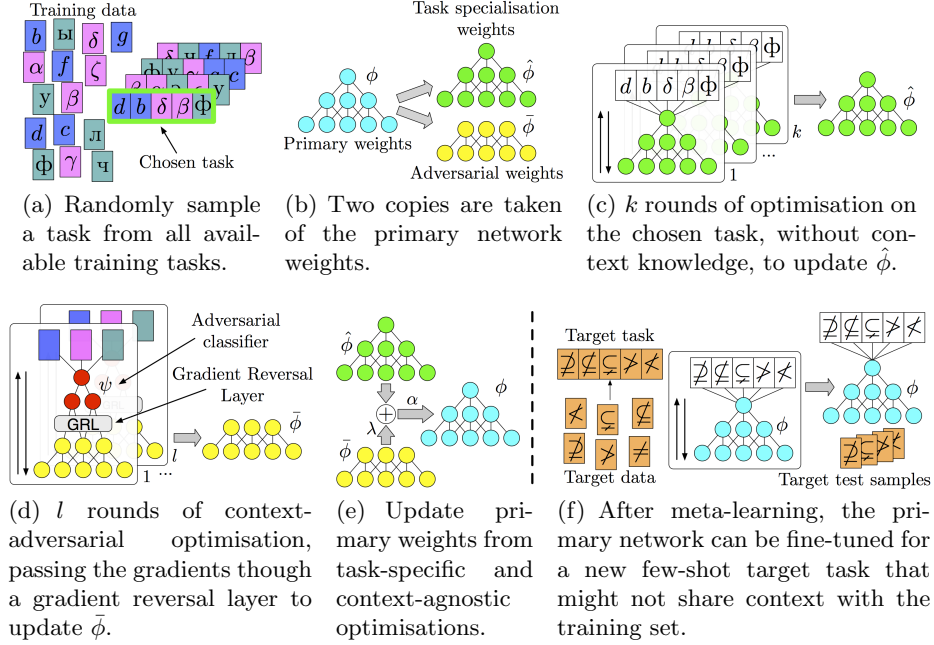


**Fig. 1.** Visualisation of how context (e.g. alphabets, shown as different colours) can contribute to train/target splits. In commonly-used split (a), a classifier could overfit on context with no ill effects. If there is novel context, as in (b), this will prove problematic. In this paper, we show how context-agnostic meta-learning can benefit performance on few-shot target tasks without shared context.

problem for meta-learners, as they can overfit on context knowledge to generate an initialisation, which affects the suitability for fine-tuning for tasks with novel contexts. Prior works on meta-learning have not sought to exploit context, even when readily available [1,2,3,4,5,6,7,8,9,10,11,12,13]. We propose a meta-learning framework to tackle both task-generalisation and context-agnostic objectives, jointly. As with standard meta-learning, we aim for trained weights that are suitable for few-shot fine-tuning to target. Note that concepts of *context* and *domain* might be incorrectly confused. Domains are typically different datasets with a significant gap, whereas context is one or more distractor signals within one dataset (e.g. font or writer for character classification), and can be either discrete or continuous.

Figure 2 presents an overview of the proposed framework, illustrated on the application of character classification. We assume that both task labels (e.g. character classification) and context labels (e.g. alphabet) are available for the training data. At each iteration of meta-learning, we randomly pick a task (Fig. 2(a)), and optimise the model’s weights for both task-generalisation (Fig. 2(c)) and context-agnosticism (Fig. 2(d)) objectives. This is achieved through keeping two copies of the model’s weights (Fig. 2(b)), one for each objective, and then updating the primary weights with a mixture of both results (Fig. 2(e)). These learnt weights are not only task-generalisable but importantly have been trained in an adversarial manner on context labels.

To demonstrate the generality of our framework, and the opportunities in considering context, we show that it is applicable to three commonly used few-shot meta-learning algorithms [1,4,7], and test our context-agnostic meta-learning framework on four diverse problems, showing clear improvements compared to prior work and baselines. The first problem (Sec 4) is Omniglot character classification [14]. We show that when using an alphabet-based split, our approach improves over non context-aware meta-learning approaches by 4.3%. The second (Sec 5) is Mini-ImageNet [10] few-shot classification, where image classification is the task, and broader class group labels are the context. An im-



**Fig. 2.** A visualisation of the proposed context-agnostic meta-learning approach through a character classification example (context shown as character colours) using an alphabet-based split (Fig. 1(b)). The method is detailed in Algorithm 1, where (a) to (e) corresponds to one outer loop iteration, which is repeated on random training tasks. (f) shows fine-tuning to target.

provement of 2.8% is observed when utilising our approach. The third (Sec 6) is few-shot classification CUB [15], where the primary colour of each bird (taken from annotations in metadata) is the context. An improvement of 1.9% is found in this case. The fourth (Sec 7) is predicting energy expenditure of people performing daily activities from video [16]. For this problem, we consider calorie prediction as the task, and the identities as the context. We show that our approach drops the Mean Square Error (MSE) from 2.0 to 1.4.

## 2 Related Work

**Few-shot Learning:** Existing few-shot methods belong to one of three categories: generative approaches [17,18], embedding-based meta-learners [9,10,11] and adaptation-based meta-learners [1,2,3,4,5,6,7,8,12,13]. Adaptation-based meta-learners produce initial models which can be fine-tuned quickly to unseen tasks, using limited labelled data. One widely-used method is Model Agnostic Meta-Learning (MAML) [1], where repeated specialisation on tasks drawn from the training set encourages the ability to adapt to new tasks with little data. Later

variations on this approach include promoting training stability [4] and improving training speed and performance on more realistic problems with deeper architectures [7]. Some works have learned alternative training curricula [3] or modified the task specialisation [2,8]. Others have learned alternative fine-tuning mechanisms [12,13] or pseudo-random labels [6] to help with adaptation to unseen tasks. These adaptation-based meta-learners contrast with embedding-based meta-learners, which find a space where the few-shot task can be embedded. A classifier is then constructed in this space, e.g. by comparing distances of target samples to seen source samples [10].

None of the above works have exploited context available from metadata of the training data. Further, they have been evaluated on datasets where additional context knowledge is not available [19,18], where context is shared between the training and target split [14,10] or combinations of the above [20,13]. We select adaptation-based meta-learning as the most suitable candidate for few-shot tasks with context. This is because there is likely to be insufficient target data for generative approaches, and target samples from a novel context are unlikely to embed well in the space constructed by embedding-based meta-learners.

**Domain Adaptation/Generalisation:** Different from domains, contexts are additional labels present within the same dataset, can be continuous and one sample could be associated with multiple contexts. However, methods that attempt domain adaptation and generalisation are relevant for achieving context-agnostic learning. Domain adaptation techniques aim to align source and target data. Some works use domain statistics to apply transformations to the feature space [21], minimise alignment errors [22], generate synthetic target data [23,24] or learn from multiple domains concurrently [25,26,27]. Adversarial domain classifiers have also been used to adapt a single [28,29,30] and multiple [31] source domains to a target domain. The disadvantage of all these approaches is that sufficient target data is required, making them unsuitable for few-shot learning. Domain generalisation works find representations agnostic to the dataset a sample is from. Approaches include regularisation [32], episodic training [33,34] and adversarial learning [35]. In this paper, we build on adversarial training, as in [28,29,30,31,35] for context-agnostic few-shot learning.

### 3 Proposed Method

We start Section 3.1 by formulating the problem, and explaining how it differs from commonly-tackled meta-learning problems. In Section 3.2, we detail our proposal to introduce context-agnostic training during meta-learning.

#### 3.1 Problem Formulation

**Commonalities to other meta-learning approaches:** The input to our method is labelled training data for a number of tasks, as well as limited (i.e. few-shot) labelled data for target tasks. Adaptation-based meta-learning is distinct from other learning approaches in that the trained model is not directly

used for inference. Instead, it is optimised for fine-tuning to a target task. These approaches have two stages: (1) the meta-learning stage - generalisable weights across tasks are learnt, suitable for fine-tuning, and (2) the fine-tuning to target stage - initialisation weights from the meta-learning stage are updated given a limited amount of labelled data from the target task. This fine-tuned model is then used for inference on test data on the target task. Throughout this section, we will focus on stage (1), i.e. the meta-learning stage, as this is where our contribution lies.

**Our novelty:** We consider problems where the unseen target task does not share context labels with the training data. We assume each training sample has both a task label and a context label. The context labels are purely auxiliary - they are not the prediction target of the main network. We utilise context labels to achieve context-agnostic meta-learning using tasks drawn from the training set and argue that incorporating context-agnosticism provides better generalisation. This is particularly important when the set of context labels in the training data is small, increasing the potential discrepancy between tasks.

### 3.2 Context-Agnostic Meta-Learning

Our contribution is applicable to adaptation-based meta-learning algorithms which are trained in an episodic manner. This means they use an inner update loop to fine-tune the network weights on a single task, and an outer update loop which incorporates changes made by the inner loop into a set of primary network weights [1,2,4,5,7]. To recap, none of these algorithms exploit context knowledge, and although they differ in the way they specialise to a single task in the inner loop, they all share a common objective:

$$\min_{\phi} \mathbb{E}_{\tau} [L_{\tau} (U_{\tau}^k (\phi))], \quad (1)$$

where  $\phi$  are the network weights,  $\tau$  is a randomly sampled task and  $L_{\tau}$  is the loss for this task.  $U_{\tau}$  denotes an update which is applied  $k$  times, using data from task  $\tau$ . Algorithm 1 shows (in black) the core of the method employed by [1,4,7], including the inner and outer loop structure common to this class of meta-learning technique. They differ in the way they calculate and backpropagate  $\nabla L_{\tau}$  in the inner specialisation loop (where different order gradients are applied, and various other training tricks are used). This step appears in Algorithm 1 L7-10 and Fig. 2(c). However, they can all be modified to become context-agnostic in the same way - this is our main contribution (shown in blue in the algorithm), which we discuss next.

To achieve context-agnostic meta-learning, we propose to train a context-adversarial network alongside the task-specialised network. This provides a second objective to our meta-learning. We update the meta-learning objective from Eq. 1 to include this context-adversarial objective, to become

$$\min_{\phi, \psi} \mathbb{E}_{\tau} [L_{\tau} (U_{\tau}^k (\phi)) + \lambda L_C (U_C^l (\psi, \phi))], \quad (2)$$

```

1 Initialise primary network with parameters  $\phi$ .
2 Initialise adversarial network with parameters  $\psi$ .
3 Link primary and adversarial networks with GRL
4 for Iteration in outer loop do
5     Select random task  $\tau$ .
6     Set  $\hat{\phi} = \phi$  and  $\bar{\phi} = \phi$ .
7     for Iteration in inner specialisation loop do
8         Construct batch with samples from task  $\tau$ .
9         Calculate  $L_\tau$ .
10        Optimise  $\hat{\phi}$  w.r.t.  $L_\tau$ .
11    end
12    for Iteration in inner adversarial loop do
13        Construct batch with samples from training dataset.
14        Add context label noise with probability  $\epsilon$ .
15        Calculate  $L_C$ .
16        Optimise  $\psi$  and  $\bar{\phi}$  w.r.t.  $L_C$ 
17    end
18    Update  $\phi \leftarrow \phi + \alpha(\hat{\phi} - \phi + \lambda(\bar{\phi} - \phi))$ .
19 end

```

**Algorithm 1:** Context-agnostic meta-learning framework. Proposed additions which can be encapsulated by existing adaptation-based meta-learning approaches, such as [1,4,7], are in blue.

where  $L_C$  is a context loss, given by an associated context network with weights  $\psi$ , which acts on the output of the network with weights  $\phi$ .  $U_C(\psi, \phi)$  is the adversarial update which is performed  $l$  times. The relative contribution of  $L_C$  is controlled by  $\lambda$ . Because  $L_C$  and  $L_\tau$  both operate on  $\phi$ , they are linked and should be optimised jointly. Equation 2 can thus be decomposed into two optimisations:

$$\phi = \arg \min_{\phi} (L_\tau (U_\tau^k(\phi)) - \lambda L_C (U_C^l(\psi, \phi))) \quad (3)$$

$$\psi = \arg \min_{\psi} (L_C (U_C^l(\psi, \phi))) \quad (4)$$

We can observe the adversarial nature of  $L_C$  in Eqs. 3 and 4, where, while  $\psi$  attempts to minimise  $L_C$ ,  $\phi$  attempts to extract features which are context-agnostic (i.e. maximise  $L_C$ ). To optimise, we proceed with two steps. The first is to update the context predictor  $\psi$  using the gradient  $\nabla_\psi L_C(\psi, \phi)$ . This is performed  $l$  times, which we write as

$$U_C^l(\nabla_\psi L_C(\psi, \phi)) \quad (5)$$

A higher  $l$  means the adversarial network trains quicker, when balanced against  $k$  to ensure  $\psi$  and  $\phi$  learn together in an efficient manner. The second step is to update the primary network with weights  $\phi$  with the gradient

$$\nabla_\phi L_\tau (U_\tau^k(\phi)) - \lambda \nabla_\phi L_C (U_C^l(\psi, \phi)) \quad (6)$$

The first term corresponds to the contribution of the task-specific inner loop. The method in [7] reduces this quantity to  $(\phi - U_\tau^k(\phi)) / \alpha$ , where  $\alpha$  is the learning rate.  $\lambda$  is a weighting factor for the contribution from the adversarial classifier, which can analogously be reduced to  $\lambda (\phi - U_C^l(\psi, \phi)) / \alpha$ . It can be incorporated by backpropagating the loss from  $\psi$  through a gradient reversal layer (GRL) to  $\phi$ . As well as performing Eqs. 5 and 6, we also perform each iteration of the  $l$  adversarial updates  $U_C$  with respect to  $\psi$  and  $\phi$  concurrently.

In practice, the process above can be simplified by taking two copies of the primary weights at the start of the process as shown in Algorithm 1, which matches the illustration in Fig. 2. At each outer iteration, we first choose a task (Algorithm 1 L5) and make two copies of the primary weights  $\phi$  (L6):  $\hat{\phi}$  (weights used for the task-specialisation inner loop) and  $\bar{\phi}$  (weights used for the context-adversarial inner loop). The task specialisation loop is then run on  $\hat{\phi}$  (L7-10). Next, the adversarial loop is run on  $\bar{\phi}$  and  $\psi$  (L12-17). The primary weights  $\phi$  are updated using weighted contributions from task-specialisation ( $\hat{\phi}$ ) and context-generalisation ( $\bar{\phi}$ ) (L18). Note that using two separate copies of the weights ensures that the task-specialisation inner loop is as similar as possible to the one fine-tuned for the target task.

The optimiser state and weights for the adversarial network with weights  $\psi$  are persistent between outer loop iterations so  $\psi$  can learn context as training progresses. This contrasts with the optimisers acting on the  $\hat{\phi}$  and  $\bar{\phi}$ , which are reset every outer loop iteration for the next randomly selected task to encourage the initialisation to be suitable for fast adaptation to a novel task.

Following standard meta-learning approaches, the weight initialisations  $\phi$  can be fine-tuned to an unseen target task. After fine-tuning on the few-shot labelled data from target tasks, this updated model can be used for inference on unlabelled data from these target tasks (see Fig. 2(f)). No context labels are required for the target, as the model is trained to be context-agnostic. Our method is thus suitable for fine-tuning to the target task when new context is encountered, as well as when contexts overlap.

Next, we explore four problems for evaluation. Recall that our approach assumes both task and context labels are available during training. In all our cases studies, we select datasets where context is available, or can be discovered, from the metadata.

## 4 Case Study 1: Character Classification

**Problem Definition.** Our first case study is few-shot image classification benchmark - Omniglot [14]. We consider the task as character classification and the context as which alphabet a character is from. We follow the standard setup introduced in [10], which consists of 1- and 5-shot learning on sets of 5 and 20 characters (5- or 20-way) from 50 alphabets. However, we make one major and important change. Recall, we have suggested that existing meta-learning techniques are not designed to handle context within the training set, or context-discrepancy between training and target. The protocol from [10] uses a *charac-*

*ter*-based split, where an alphabet can contribute characters to *both* train and target tasks (Fig. 1(a)). Instead, we eliminate this overlap by ensuring that the characters are from different alphabets, i.e. an *alphabet*-based split (Fig. 1(b)).

**Evaluation and Baselines.** We evaluate the proposed context-agnostic framework using three meta-learners: MAML++ [4], MAML [1] and REPTILE [7]. Note that other adaptation-based meta-learning methods could also be used by substituting in their specific inner-specialisation loops [2,5]. Unmodified versions are used as baselines, and are compared against versions which are modified with our proposed context agnostic (CA) component. We accordingly refer to our modified algorithms as CA-MAML++, CA-MAML and CA-REPTILE. We report results without transduction, that is batch normalisation statistics are not calculated from the entire target set in advance of individual sample classification. This is more representative of a practical application. As in [10], the metric is top-1 character classification accuracy. We run experiments on the full dataset, and also on a reduced number of alphabets. With 5 alphabets, for example, characters from 4 alphabets are used for training, and a few-shot task is chosen from the 5th alphabet only. As the number of alphabets in training decreases, a larger context gap would be expected between training and target. We report averages over 10 random train/target splits, and keep these splits consistent between experiments on the same number of alphabets.

**Implementation Details.** The widely-used architecture, optimiser and hyperparameters introduced in [10], are used. We implement the adversarial context predictor in the proposed context-agnostic methods as a single layer which takes the penultimate features layer (256D) as input with a cross-entropy loss applied to the output, predicting the alphabet. Context label randomisation is used in the adversarial classifier, where 20% of the context labels are changed. This stops the context adversarial loss tending to zero too quickly (similar to label smoothing [36]). We use  $l = 3$  (Eq. 2) for all Omniglot experiments. The context-agnostic component increases the training time by 20% for all methods.

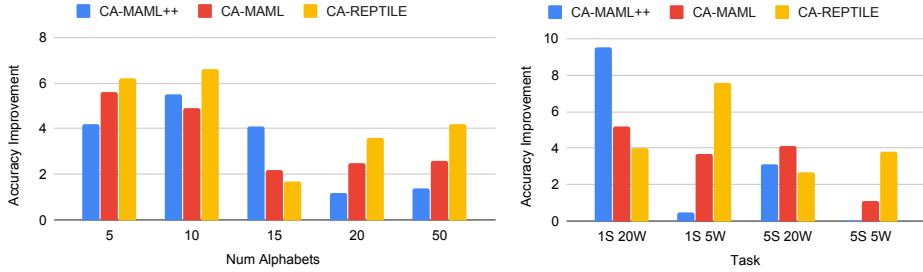
**Results.** Table 1 shows the results of the proposed framework applied to [4,1,7] on 5-50 alphabets, using the alphabet-based split shown in Fig. 1(b). We report results per method, to show our proposed context-agnostic component improves on average across all methods, tasks and numbers of alphabets. 85% of individual method/task/alphabet combinations show an improvement, with a further 10% being comparable (within 1% accuracy). Overall, the proposed framework gives an average performance increase of 4.3%. This improvement is most pronounced for smaller numbers of alphabets (e.g. average improvements of  $\geq 6.2\%$ , 4.9% and 4.2% for 5 and 10 alphabets for [7,1,4] respectively). This trend is shown in Fig. 3(a), and supports our earlier hypothesis that the inclusion of a context-agnostic component is most beneficial when the context overlap between the train and target data is smaller. Fig. 3(b) shows the improvement for each XS YW task, averaged over the number of alphabets. Larger improvements are observed for all methods on the 1-shot versions of 5- and 20-way tasks, with [7] improving the most on 1S 5W and [1,4] improving the most on 1S 20W.



**Table 1.** Character classification accuracy on Omniglot, using an alphabet-based split, with the number of training alphabets varied between 5 and 50. XS YW indicates X-shot fine-tuning at a Y-way classification tasks. Base methods are compared against context-agnostic (CA) versions.

Task	Method	Number of Alphabets				
		5	10	15	20	50
1S 20W	MAML++ [4]	58.7	57.2	64.7	<b>85.6</b>	89.6
	CA-MAML++	<b>72.3</b>	<b>67.6</b>	<b>82.4</b>	84.8	<b>90.9</b>
	MAML [1]	61.4	78.2	81.5	83.7	87.5
	CA-MAML	<b>69.8</b>	<b>82.8</b>	<b>82.1</b>	<b>89.8</b>	<b>93.8</b>
	REPTILE [7]	11.9	18.1	37.6	51.6	64.9
	CA-REPTILE	<b>20.7</b>	<b>21.8</b>	<b>39.5</b>	<b>55.5</b>	<b>66.5</b>
1S 5W	MAML++ [4]	97.4	96.2	<b>94.9</b>	93.4	93.7
	CA-MAML++	<b>98.1</b>	<b>97.1</b>	90.1	<b>95.8</b>	<b>97.1</b>
	MAML [1]	86.1	87.0	<b>96.1</b>	94.4	90.5
	CA-MAML	<b>94.5</b>	<b>91.3</b>	94.7	<b>96.0</b>	<b>96.2</b>
	REPTILE [7]	52.2	68.8	79.4	75.5	77.5
	CA-REPTILE	<b>62.2</b>	<b>76.9</b>	<b>83.4</b>	<b>83.2</b>	<b>85.5</b>

Task	Method	Number of Alphabets				
		5	10	15	20	50
5S 20W	MAML++ [4]	81.0	84.1	92.4	93.5	95.8
	CA-MAML++	<b>84.8</b>	<b>90.8</b>	<b>96.0</b>	<b>94.5</b>	<b>96.3</b>
	MAML [1]	81.7	83.8	84.0	91.2	<b>89.0</b>
	CA-MAML	<b>86.0</b>	<b>91.8</b>	<b>92.9</b>	<b>93.1</b>	86.9
5S 5W	REPTILE [7]	58.4	68.1	76.7	<b>76.0</b>	78.0
	CA-REPTILE	<b>61.1</b>	<b>73.7</b>	<b>78.3</b>	75.8	<b>81.6</b>
5S 5W	MAML++ [4]	<b>99.4</b>	<b>99.3</b>	<b>98.7</b>	97.0	96.8
	CA-MAML++	99.3	98.6	98.5	<b>99.4</b>	<b>96.9</b>
	MAML [1]	96.6	95.8	97.2	97.9	98.9
	CA-MAML	<b>97.8</b>	<b>98.5</b>	<b>97.6</b>	<b>98.6</b>	<b>99.1</b>
	REPTILE [7]	85.2	85.6	<b>93.2</b>	88.5	89.4
	CA-REPTILE	<b>88.3</b>	<b>94.4</b>	92.4	<b>91.6</b>	<b>92.9</b>

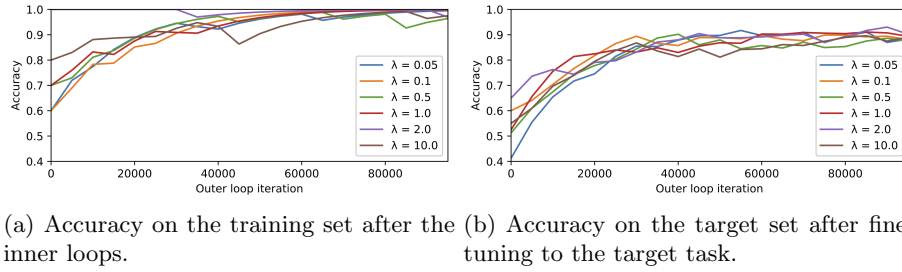


(a) Averaged over the 1- and 5-shot, 5- and 20-way tasks, showing the effect of the number of unique context labels (i.e. alphabets). (b) Averaged over number of alphabets (5, 10, 15, 20 and 50), showing how each task is affected.

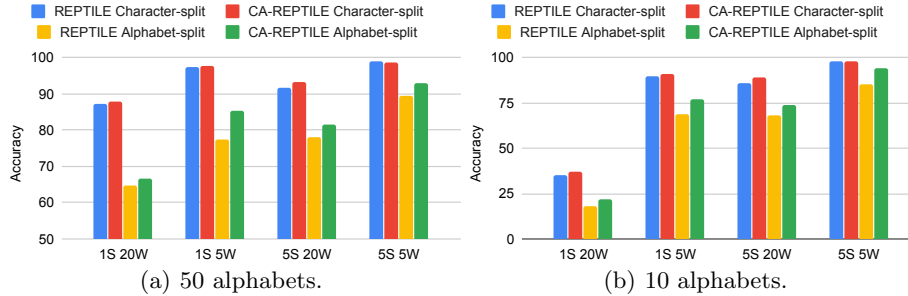
**Fig. 3.** Accuracy improvements given by our context-agnostic (CA-) versions of [4,1,7] using the alphabet-based split (shown in Fig. 1(b)).

For the ablation studies, we use [7] as our base meta-learner as it is the least computationally expensive. Based on preliminary studies, we believe the behaviour is consistent, and the conclusions stand, for the other methods. In the results above, we used  $\lambda = 1.0$  for the contribution of our adversarial component  $\lambda$  (Eq. 1). Next, we provide results on how varying  $\lambda$  can affect the model's performance. For this, we use 5S 5W, 10 alphabet task. Fig. 4 shows training progress with  $\lambda = \{10.0, 2.0, 1.0, 0.5, 0.1\}$ . We can see that a high weighting ( $\lambda = 10.0$ ) causes a drop in training accuracy around iteration 40K, as the optimisation prioritises becoming context-agnostic over the ability to specialise to a task. However, the figure shows reasonable robustness to the choice of  $\lambda$ .

Next, we investigate the differences between character-based and alphabet-based training/target splits (visualised in Fig. 1). Fig. 5 shows the effects of context-agnosticism when evaluating on character-based splits and alphabet-



**Fig. 4.** These plots show how the weighting ( $\lambda$ ) of the context-adversarial component affects training and target performance during one run of the 5-shot/5-way 10 alphabet task using an alphabet-based split.



**Fig. 5.** Comparison of character-based and alphabet-based training/target splits using 50 and 10 alphabets.

based splits. Fig. 5(a) uses 50 alphabets for comparison, and Fig. 5(b) uses 10 alphabets. While both approaches are comparable on character-based splits (blue vs red), we show a clear improvement in using our context-agnostic meta-learning approach when tested on alphabet-based splits (yellow vs green). This is a sterner test due to the training and target sets being made up from data with different contexts. The context-agnostic version is significantly better for all cases and both alphabet sizes.

Finally, as previous approaches only evaluate on the easier character-based split for Omniglot, using all 50 alphabets, we provide comparative results to published works on this setup. We list reported results from [1,4,7] as well as our replications to ensure a direct comparison (the same codebase and splits can be used with and without the context-agnostic component). For this setup, we use the same data augmentation as [1,4,7]. Results are given in Table 2, which confirms that context-agnostic versions of the base methods achieve comparable performance, despite there being shared context between source and target.

In summary, this section presented experiments on the Omniglot character classification dataset. We show that, on average, our proposed context-agnostic approach gives performance improvements across all methods and tasks, partic-

**Table 2.** Comparative results on Omniglot using the standard character-based split. \*: results reported in cited papers. Even though both training and target tasks share context, our CA contribution maintains performance on this standard split.

Method	5S 5W	1S 5W	5S 20W	1S 20W
MAML++ [4]*	99.9	99.4	99.3	97.7
MAML++ [4]	99.9	99.5	98.7	95.4
CA-MAML++	99.8	99.5	98.8	95.6
MAML [1]*	99.8	98.6	98.9	95.8
MAML [1]	99.8	99.3	97.0	92.3
CA-MAML	99.8	99.3	97.2	94.8
REPTILE [7]*	98.9	95.4	96.7	88.1
REPTILE [7]	98.9	97.3	96.4	87.3
CA-REPTILE	98.6	97.6	95.9	87.8

ularly for smaller alphabet sizes, which introduce a bigger context gap between training and target.

## 5 Case Study 2: General Image Classification

**Problem Definition.** Our second case study uses the few-shot image classification benchmark - Mini-ImageNet [10]. We use the experimental setup introduced in [10], where the task is a 1- or 5-shot 5-way classification problem. Similar to our previous case study, we aim for context labels, and a context-based split. This dataset has no readily-available context labels, and there is a large overlap between the train and target splits (e.g. 3 breeds of dog in target, 12 in train). We address this by manually assigning 12 superclass labels, which we use as context. We then ensure that superclasses used for training and testing are distinct.

**Evaluation, Baselines and Implementation.** Similar to Section 4, we evaluate using MAML++ [4] and MAML [1]. Unmodified versions are used as baselines, and are compared against versions which are modified with our proposed CA component. Transduction is not used, and the metric is top-1 image classification accuracy. The same architecture, hyperparameters etc. as in [4] are used. We use  $k = 5$  (Eq. 1) and  $l = 2$  (Eq. 2). Results are given for the original Mini-ImageNet splits and our superclass-based splits with context labels.

**Results.** Table 3 shows the results on the original train/target split and the new splits with no shared context. Results show comparable performance for the original split, but importantly improved performance in the context-based split. Our context-agnostic component improves over [1] and [4] by an average 3.3% on the most difficult 1S 5W task. An average 2.2% improvement is also seen on the easier 5S 5W task. Similar to Omniglot, note that few shot classification on Mini-ImageNet is more challenging (by an average of 8.7% across all methods) when there is no shared context between training and target data.

**Table 3.** Results on Mini-ImageNet and CUB using the original splits which have shared context between train and target tasks, and the new context-based splits with no shared context between training and target tasks.

Method	Mini-ImageNet				CUB			
	Original split		Context Split		Original split		Context Split	
	1S 5W	5S 5W	1S 5W	5S 5W	1S 5W	5S 5W	1S 5W	5S 5W
MAML++ [4]	<b>52.0</b>	<b>68.1</b>	40.1	60.1	<b>38.7</b>	57.2	42.2	56.7
CA-MAML++	51.8	<b>68.1</b>	<b>44.4</b>	<b>61.5</b>	38.0	<b>58.4</b>	<b>43.3</b>	<b>57.9</b>
MAML [1]	<b>48.3</b>	<b>64.3</b>	41.1	56.5	42.5	<b>56.1</b>	37.7	54.7
CA-MAML	<b>48.3</b>	64.2	<b>43.3</b>	<b>59.5</b>	<b>42.6</b>	55.9	<b>40.3</b>	<b>57.5</b>

## 6 Case Study 3: Fine-Grained Bird Classification

**Problem Definition.** For our third case study, we use the few-shot fine-grained bird classification benchmark CUB [15]. CUB contains a large amount of meta-data from human annotators. For context labels, we have taken each bird’s primary colour, but could have chosen a number of others e.g. bill shape. The CUB dataset has 200 classes, with 9 different primary colours. We ensure splits are distinct with respect to this property.

**Evaluation, Baselines and Implementation.** We use the same setup as for Mini-ImageNet (Section 5).

**Results.** Table 3 shows the results on the original train/target splits and the new splits with no shared context (i.e. no shared primary colour). When there is less less shared context between train and target data, our context-agnostic component improves over [1] and [4] by an average of 1.9% across all tasks, whilst performance is maintained on the original split.

## 7 Case Study 4: Calorie Estimation from Video

**Problem Definition.** In this fourth problem, we use the dataset from [37], where the task is to estimate energy expenditure for an input video sequence of an individual carrying out a variety of actions. Different from the first three case studies, this is a regression task, rather than a classification one, as calorie readings are continuous. The target task is to estimate the calorimeter reading for seen, as well as unseen, actions. Importantly, the individual captured forms the context. Alternative context labels could include, for example, age or Body Mass Index (BMI). Our objective is thus to perform meta-learning to generalise across actions, as well as being individual-agnostic, for calorie prediction of a new individual. We use silhouette footage and calorimeter readings from 10 participants performing a number of daily living tasks as derived from the SPHERE Calorie dataset of [16]. Using a relatively small amount of data to fine-tune to target is appropriate because collecting data from individuals using a calorimeter is expensive and cumbersome.

**Evaluation and Baselines.** Ten-fold leave-one-person-out cross-validation is used for evaluation. We report results using MSE across all videos for each

**Table 4.** MSE for all 10 participants on the Calorie dataset, using leave-one-out cross-validation. A lower MSE indicates better results. Methods with only an average reported are results taken from the referenced publications.

Method	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Avg
MET Lookup [16]	-	-	-	-	-	-	-	-	-	-	2.25
Tao et al. [37]	-	-	-	-	-	-	-	-	-	-	1.69
Pre-train only	1.21	<b>0.89</b>	0.88	1.86	1.24	<b>2.46</b>	7.50	<b>0.89</b>	1.25	3.11	2.13
Pre-train/fine-tune	0.58	1.64	0.75	0.53	1.13	4.26	5.83	1.29	1.41	3.53	2.10
REPTILE [7]	0.48	1.65	0.52	0.90	2.12	3.28	6.48	1.26	<b>0.83</b>	2.58	2.01
CA-REPTILE	<b>0.39</b>	1.11	<b>0.46</b>	<b>0.48</b>	<b>0.87</b>	2.68	<b>3.75</b>	1.07	0.87	<b>2.32</b>	<b>1.40</b>

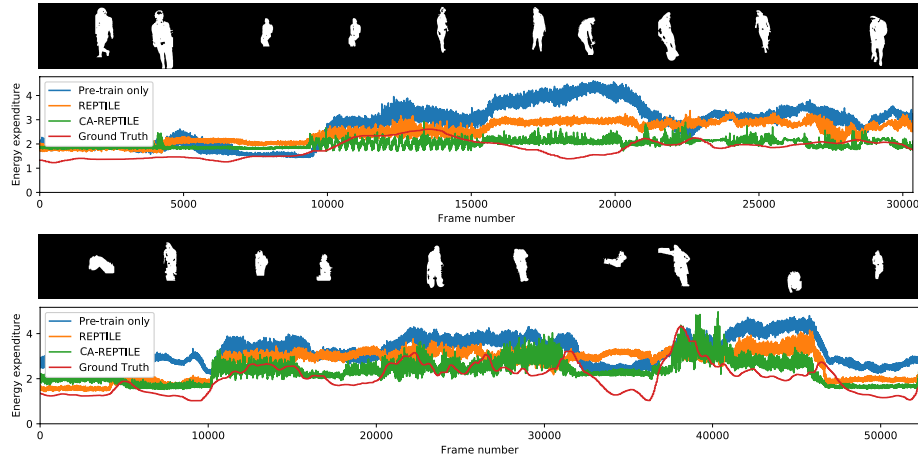
subject. For fine-tuning to target, we use labelled calorie measurements from the first 32 seconds (i.e. the first 60 video samples, where each sample is 30 frames subsampled at 1fps) of the target subject. Evaluation is then performed using the remaining data from the target subject, which is 28 minutes on average. We compare the following methods, using cross-fold, leave-one-person-out validation:

- Metabolic Equivalent (MET) from [16]. This offers a baseline of calorie estimation through a look-up table of actions and their duration. This has been used as a baseline on this dataset previously.
- Method from Tao et al. [37] that utilises IMU and depth information not used by our method.
- Pre-train - standard training process, trained on 9 subjects and tested on target subject without fine-tuning.
- Pre-train/fine-tune - standard training process on 9 subjects and fine-tuned on the target subject.
- REPTILE - meta-learning from [7] on 9 subjects and fine-tuned on target.
- CA-REPTILE - our proposed context-agnostic meta-learning approach.

Note that we chose to use [7] as the baseline few-shot method because it is less computationally expensive (important when scaling up the few shot-problem to video) than [1,4], as discussed in Section 2.

**Implementation Details.** Images are resized to 224x224, and fed to a ResNet-18 architecture [38]. No previous works have addressed this individual-agnostic personalisation problem. Following [16], it is believed that a window of 30s is required as input for energy expenditure prediction. We sample the data at 1fps and use the ResNet CNN’s output from the penultimate layer as input to a Temporal Convolutional Network (TCN) [39] for temporal reasoning. Our model is trained end-to-end using Adam [40] and contains 11.2M parameters. We use  $k = 10$  (Eq. 1) and  $l = 1$  (Eq. 2) for all Calorie experiments. A lower value of  $l$  is required than for Omniglot, as context information is easier for the adversarial network to learn (i.e. people are easier to distinguish than alphabets). MSE is used as the regression loss function. Augmentation during training consists of random crops and random rotations up to  $30^\circ$ . The same architecture is used for all baselines (except MET and [37]), making results directly comparable.

**Results.** Table 4 compares the various methods. The context-agnostic meta-learning method obtains a 35% reduction in MSE over the pre-training only, a



**Fig. 6.** Example energy expenditure predictions on two sequences from different participants in the Calorie dataset.

33% reduction over the pre-train/fine-tune model, and a 30% improvement over the non context-agnostic version. For 3 out of 10 individuals, pre-training outperforms any fine-tuning. We believe this is due to these participants performing actions at the start of the sequence in a different manner to those later. However, our context-agnostic approach offers the best fine-tuned results.

Fig. 6 shows qualitative silhouette sequences with calorimeter readings as groundtruth, which are to be compared to predictions from our method and baselines. Results demonstrate that the context-agnostic version estimates the ground truth curve better than other methods from participants with low and high energy expenditure variability.

## 8 Conclusion

In this paper, we proposed context-agnostic meta-learning that learns a network initialisation which can be fine-tuned quickly to new few-shot target problems. An adversarial context network acts on the initialisation in the meta-learning stage, along with task-specialised weights, to learn context-agnostic features capable of adapting to tasks which do not share context with the training set. This overcomes a significant drawback with current few-shot meta-learning approaches, that do not exploit context which is often readily available. The framework is evaluated on the Omniglot few-shot character classification dataset and the Mini-ImageNet and CUB few-shot image recognition tasks, where it demonstrates consistent improvements when exploiting context information. We also evaluate on a few-shot regression problem, for calorie estimation from video, showing significant improvements.

This is the first work to demonstrate the importance and potential of incorporating context into few-shot methods. We hope this would trigger follow-up works on other problems, methods and contexts.

**Data Statement:** Our work uses publicly available datasets. Proposed context-based splits are available at [github.com/tobyperrett/context\\_splits](https://github.com/tobyperrett/context_splits).

**Acknowledgement:** This work was performed under the SPHERE Next Steps Project, funded by EPSRC grant EP/R005273/1.

## References

1. Finn, C., Abbeel, P., Levine, S.: Model-Agnostic Meta-Mearning for Fast Adaptation of Deep Networks. In: International Conference on Machine Learning. (2017)
2. Rusu, A.A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., Hadsell, R.: Meta-Learning with Latent Embedding Optimization. International Conference on Learning Representations (2019)
3. Sun, Q., Chua, Y.L.T.s.: Meta-Transfer Learning for Few-Shot Learning. In: Computer Vision and Pattern Recognition. (2019)
4. Antoniou, A., Edwards, H., Storkey, A.: How to Train Your MAML. In: International Conference on Learning Representations. (2019)
5. Finn, C., Xu, K., Levine, S.: Probabilistic Model-Agnostic Meta-Learning. In: Advances in Neural Information Processing Systems. (2018)
6. Sun, Q., Li, X., Liu, Y., Zheng, S., Chua, T.S., Schiele, B.: Learning to Self-Train for Semi-Supervised Few-Shot Classification. In: Advances in Neural Information Processing Systems. (2019)
7. Nichol, A., Achiam, J., Schulman, J.: On First-Order Meta-Learning Algorithms. arXiv 1803.02999 (2018)
8. Bertinetto, L., Henriques, J.F., Torr, P.H.S., Vedaldi, A.: Meta-learning with Differentiable Closed-Form Solvers. International Conference on Learning Representations (2019)
9. Snell, J., Swersky, K., Zemel, R.: Prototypical Networks for Few-Shot Learning. Advances in Neural Information Processing Systems (2017)
10. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching Networks for One Shot Learning. In: Advances in Neural Information Processing Systems. (2016)
11. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-Learning for Semi-Supervised Few-Shot Classification. In: International Conference on Learning Representations. (2018)
12. Requeima, J., Gordon, J., Bronskill, J., Nowozin, S., Turner, R.E.: Fast and Flexible Multi-Task Classification Using Conditional Neural Adaptive Processes. In: Advances in Neural Information Processing Systems. (2019)
13. Tseng, H.Y., Lee, H.Y., Huang, J.B., Yang, M.H.: Cross-domain few-shot classification via learned feature-wise transformation. In: International Conference on Learning Representations. (2020)
14. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. *Science* **350** (2015) 1332–1338
15. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Technical report (2011)
16. Tao, L., Burghardt, T., Mirmehdi, M., Damen, D., Cooper, A., Hannuna, S., Campani, M., Paiement, A., Craddock, I.: Calorie counter: RGB-Depth Visual Estimation of Energy Expenditure at Home. In: Asian Conference on Computer Vision Workshops. (2016)

17. Zhang, R., Che, T., Bengio, Y., Ghahramani, Z., Song, Y.: Metagan: An Adversarial Approach to Few-Shot Learning. *Advances in Neural Information Processing Systems* (2018)
18. Dwivedi, S.K., Gupta, V., Mitra, R., Ahmed, S., Jain, A.: ProtoGAN: Towards Few Shot Learning for Action Recognition. In: *Computer Vision and Pattern Recognition*. (2019)
19. Oreshkin, B.N., Rodriguez, P., Lacoste, A.: Tadam: Task Dependent Adaptive Metric for Improved Few-Shot Learning. In: *Advances in Neural Information Processing Systems*. (2018)
20. Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.A., Larochelle, H.: Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. In: *International Conference on Learning Representations*. (2019)
21. Panareda Busto, Pau and Gall, J.: Open Set Domain Adaptation. In: *International Conference on Computer Vision*. (2017)
22. Haeusser, P., Frerix, T., Mordvintsev, A., Cremers, D.: Associative Domain Adaptation. In: *International Conference on Computer Vision*. (2017)
23. Hoffman, J., Tzeng, E., Park, T., Phillip, J.y.Z., Kate, I., Alexei, S., Darrell, T.: CyCADA : Cycle-Consistent Adversarial Domain Adaptation. In: *International Conference on Machine Learning*. (2018)
24. Huang, S.W., Lin, C.T., Chen, S.P., Wu, Y.Y., Hsu, P.H., Lai, S.H.: AugGAN: Cross Domain Adaptation with GAN-Based Data Augmentation. In: *European Conference on Computer Vision*. (2018)
25. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning Multiple Misual Domains with Residual Adapters. *Advances in Neural Information Processing Systems* (2017)
26. Perrett, T., Damen, D.: DDLSTM: Dual-Domain LSTM for Cross-Dataset Action Recognition. In: *Computer Vision and Pattern Recognition*. (2019)
27. Li, Y., Vasconcelos, N.: Efficient Multi-Domain Network Learning by Covariance Normalization. In: *Computer Vision and Pattern Recognition*. (2019)
28. Ganin, Y., Lempitsky, V.: Unsupervised Domain Adaptation by Backpropagation. In: *International Conference on Machine Learning*. (2015)
29. Zhang, Y., Tang, H., Jia, K., Tan, M.: Domain-Symmetric Networks for Adversarial Domain Adaptation. In: *Computer Vision and Pattern Recognition*. (2019)
30. Kang, B., Feng, J.: Transferable Meta Learning Across Domains. In: *Conference on Uncertainty in Artificial Intelligence*. (2018)
31. Schoenauer-Sebag, Alice and Heinrich, Louise and Schoenauer, Marc and Sebag, Michele and Wu, Lani F and Altschuler, S.J.: Multi-Domain Adversarial Learning. In: *International Conference on Learning Representations*. (2019)
32. Balaji, Y., Sankaranarayanan, S., Chellappa, R.: Metareg: Towards domain generalization using meta-regularization. In: *Advances in Neural Information Processing Systems*. (2018)
33. Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.Z., Hospedales, T.: Episodic training for domain generalization. In: *International Conference on Computer Vision*. (2019)
34. Dou, Q., Castro, D.C., Kamnitsas, K., Glocker, B.: Domain Generalization via Model-Agnostic Learning of Semantic Features. In: *Advances in Neural Information Processing Systems*. (2019)
35. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain Generalization with Adversarial Feature Learning. In: *Computer Vision and Pattern Recognition*. (2018)



36. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved Techniques for Training GANs. In: *Advances in Neural Information Processing Systems*. (2016)
37. Tao, L., Burghardt, T., Mirmehdi, M., Damen, D., Cooper, A., Camplani, M., Hannuna, S., Paiement, A., Craddock, I.: Energy Expenditure Estimation Using Visual and Inertial Sensors. *IET Computer Vision* **12** (2018) 36–47
38. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition* (2016)
39. Bai, S., Kolter, J.Z., Koltun, V.: An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv 1803.01271* (2018)
40. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations*. (2015)