Supplementary Materials: Incremental Meta-Learning via Episodic Replay Distillation for Few-Shot Image Recognition

Anonymous CVPR submission

Paper ID 11

1. Comparison between EIML and ERD

The proposed distillation strategy, proposed in this paper, more efficiently exploits the exemplars for knowledge transfer than the ERD method proposed [1]. ERD performs meta distillation over all classes by exemplar replay, while EIML (the variant of ERD that uses exemplars [1]) only distills the new classes with the old class prototypes. Since the old classes are drifting due to forgetting in EIML and distillation is more efficient when data used for distillation is more similar to previous task data, aligning the new classes (which can be far away in embedding space) to old prototypes further impedes the learning of the meta model. In ERD, however, we also use the exemplars to distill the knowledge (defined by a few-shot problem on both old and new classes). We argue (and will later experimentally verify) that this leads to more efficient knowledge transfer.

We compare between EIML and ERD (with only the cross-task sub-episode) in Figure 7. As a summarization, EIML is under-utilizing exemplars only using them to compute the prototype means. ERD also uses the exemplars during the meta-learning as well as for improved distillation. Moreover, we further enhance our ERD model with extra exemplar sub-episodes to improve the results, which is absent in EIML.

In addition, a full scheme of our method ERD based on ProtoNets is shown in Algorithm. 1.

Algorithm 1: Episodic Replay Distillation based on ProtoNets

Input : New task data X_t, exemplar memory E_{t-1}, old model $f_{\theta_{t-1}}$.
Output: New model f_{θ_t}, new exemplar set E_t

Initialize new model: $f_{\theta_t} \leftarrow f_{\theta_{t-1}}$.
Sample exemplar sub-episodes from E_{t-1} and cross-task episodes from X_t and E_{t-1} (Figure 2a).
for *all sampled cross-task and exemplar sub-episodes* **do**
    Compute L_{meta} (Eq. 5) and L^m_{dist} (Eq.7) from cross-task sub-episode.
    Compute L^e_{dist} (Eq. 6) from exemplar sub-episode.
    Update f_{θ_t} using final loss: $L = L_{\text{meta}} + \lambda_m L^m_{\text{dist}} + \lambda_e L^e_{dist}$ (Eq. 8)
end
Select new exemplars from X_t using nearest to center criterion and merge them with E_{t-1} to form E_t.
while $|E_t| > M$ **do**
    Randomly remove exemplars from the class with the most.
end
return f_{θ_t}, E_t

2. Extra ablation studies

Ablation on distillation and cross-task meta losses. We start by ablating the two main novel contributions of our paper: the use of exemplars in the cross-task meta-loss and our proposed knowledge distillation method. This ablation study is performed on CIFAR100 in the 16-task 1-shot/5-way scenario with 4-Conv as the backbone. We focus on *meta-test accuracy* to compare among variants since this is the most important evaluation metric in incremental meta-learning.

In Table 2 we vary the way the distillation and cross-task meta losses are computed. In EIML, the meta loss is computed only with new classes and the distillation loss aligns new classes with old classes. In ERD the meta loss is computed over all classes and the distillation loss aligns all classes. From the comparison among these five variants, we observe that both components contribute to the final performance gain with respect to EIML. The underlying reason is that EIML ignores the importance of mixing old and new classes at incremental meta-training time. In EIML, exemplars are only used to compute *previous* task prototypes, which under-exploits their potential. Furthermore, a proper distillation loss over exemplars is very important. We incrementally construct the classifier using two sub-episodes:

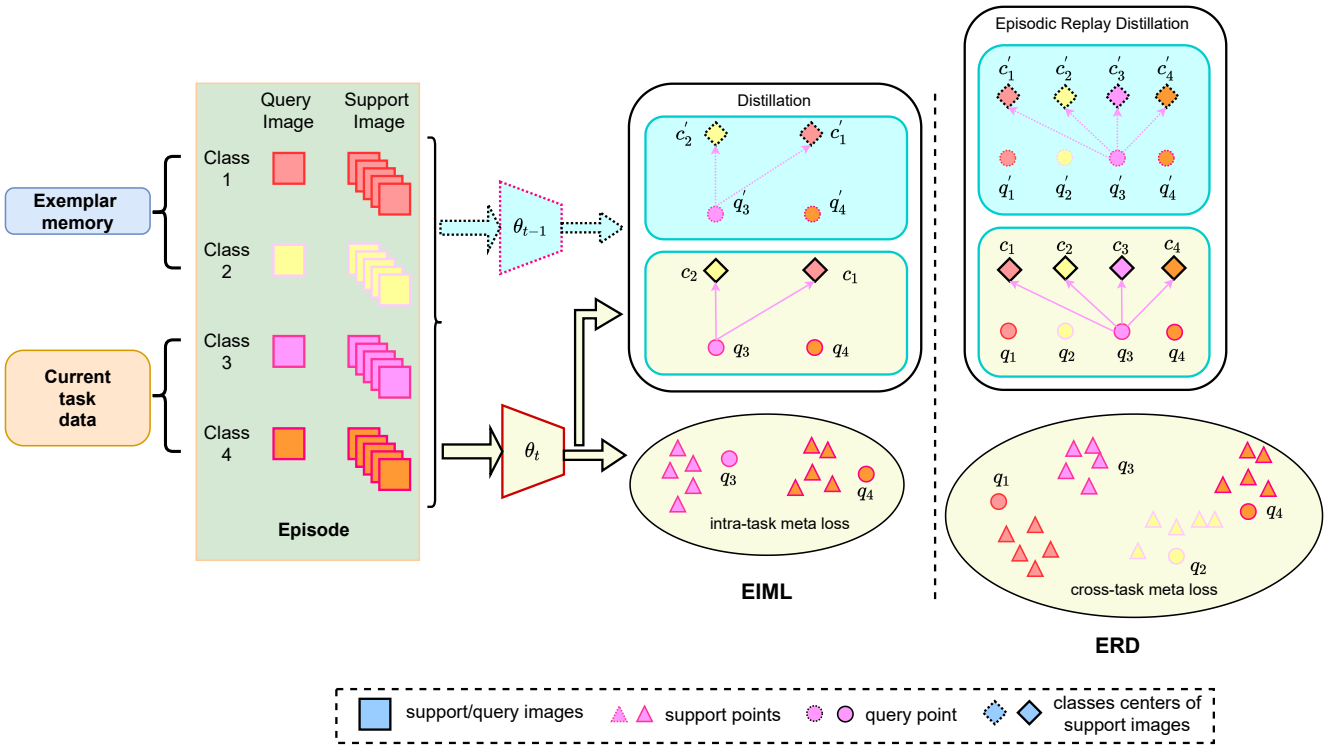


Figure 7. Illustration of novelty of our method ERD with respect to previous state-of-the-art method EIML [1]. We sample an episode that is passed through the previous model θ_{t-1} and current model θ_t. EIML only uses exemplars to compute prototypes used during distillation. While, in ERD, we additionally use exemplars to improve meta-learning by using a cross-task meta-learning loss, and we also apply them in our improved knowledge distillation approach called Episodic Replay Distillation.

the exemplar sub-episode ensures good performance on old classes, and the cross-task sub-episode ensures discrimination of all classes. The ablation results show that the use of both sub-episodes during distillation helps to significantly improve the results on unseen classes.

Ablation on λ_m and λ_e with $P = 0.2$. In the main paper, we have shown the ablation study with $\lambda_m = 0.5$, $\lambda_e = 0$ and $\lambda_m = 0$, $\lambda_e = 0.5$ to prove the effectiveness of each loss component. In Figure 8a, we extend the ablation with $\lambda_e = \lambda_m = \{0.0, 0.1, 0.5, 1.0, 2.0, 10.0\}$. We can observe that this hyperparameter gets the best results for $\{0.5,1.0,2.0\}$.

Ablation on exemplar selection strategies. To save exemplars after each training session, we need to choose N_{ex} for each class. We performed an ablation to compare nearest-to-center, random selection, herding [2], and a simplified version of Rainbow Memory [3] (we call RB^*) in Figure 8b with $P = 0.2, \lambda = \lambda_m = \lambda_e = 0.5$. For RB^*. Since we have no joint classifier over all classes, which is needed for Rainbow Memory, we imitate the idea by selecting exemplars near the boundary or exemplars near the center of the class with equal probability. This obtains slightly better results, but essentially the exemplar selection strategies differ little in final performance.

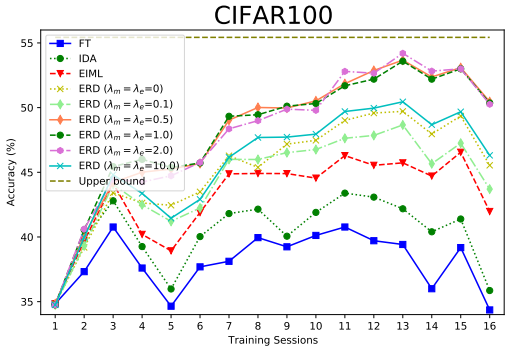
3. Results on short task sequences

We also compare our method with others on short sequences. In Tables 5 and 6, we first evaluate our model with a 4-Conv backbone on 1-shot/5-shot 5-way few-shot on three datasets. We see that FT suffers from catastrophic forgetting, and that meta-test accuracy drops dramatically and exhibits overfitting to the current task. IDA is not able to improve meta-test accuracy on Mini-ImageNet, but improves performance on CIFAR100 and CUB. As for EIML, with exemplars it shows large improvement compared to IDA. However, our method ERD outperforms EIML by a large margin after learning all four tasks. These results further confirm the observations on the 16-task setting. ERD not only achieves the best performance with less forgetting, but also gets closer to the upper bound after the last task. Note also that CIFAR100 and Mini-ImageNet are coarse-grained datasets, compared to CUB, which makes few shot classification much harder due to intra-class variability.

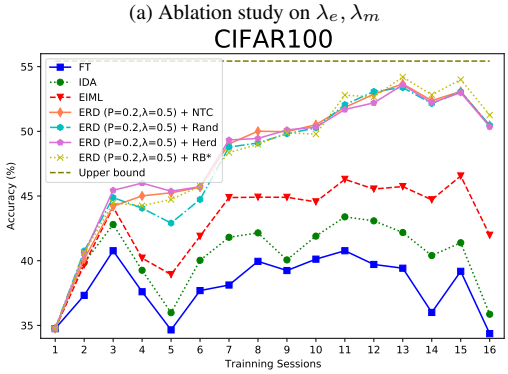
Also, we consider ResNet-12 as a backbone to show that ERD can be applied to different network architectures. Our method achieves consistently better performance over oth-

Datasets: CIFAR100, Learner: ProtoNets, Backbone: 4-Conv									
Exemplar usage				Method	Training sessions				
Meta loss	Distillation data	Old Prototype	Exemplar SE		2	4	8	16	avg
No	No	Yes	No	EIML	39.7	40.2	44.9	42.0	43.1
No	Yes	Yes	No	-	38.9	40.6	43.3	43.7	43.3
Yes	No	Yes	No	-	39.2	42.6	45.4	45.5	45.2
Yes	Yes	Yes	No	ERD	39.6	43.6	49.0	47.6	46.8
Yes	Yes	Yes	Yes	ERD	**40.1**	**45.0**	**50.0**	**50.5**	**48.1**

Table 2. Meta-test accuracy under the 1-shot/5-way 16-task setting. Here we ablate the meta loss and distillation loss. *Exemplar SE* is brief for *Exemplar sub-episode*.



(a) Ablation study on λ_e, λ_m



(b) Ablation on exemplar selection strategy

Figure 8. Ablation study on 16-task 1-shot/5-way setup on CIFAR100 with 4-Conv. We plot the meta-test accuracy to compare.

ers with much higher accuracy than using the 4-Conv backbone. Finally, in Table 7, we extend the Table 1 in the main paper with 4-task setting based on Relation Networks.

4. Task split

Table 4 is an illustration of the 16-task setting data split.

5. Confidence intervals

In Table 3 we give the meta-test accuracy with confidence intervals on Mini-ImageNet for the 1-shot/5-way/4-task scenario. The confidence intervals are relatively small with respect to average accuracy and almost the same for different methods. Therefore, it is fair to compare different

Learner	ProtoNets			
Backbone	4-Conv			
Datasets	Mini-ImageNet			
	1-shot 5-way 4-task setting			
	Upper bound: 53.24 ±0.22			
Sessions	1	2	3	4
FT	43.79 ±0.18	44.05 ±0.19	42.44 ±0.21	37.91 ±0.20
IDA	43.79 ±0.18	48.25 ±0.20	47.16 ±0.24	42.33 ±0.23
EIML	43.79 ±0.18	48.79 ±0.19	49.37 ±0.20	47.53 ±0.20
ERD	43.79 ±0.18	51.14 ±0.20	52.27 ±0.21	52.98 ±0.21

Table 3. Meta-test accuracy with confidence intervals under the 1-shot/5-way 4-task setting.

	IML training tasks				Meta-test Images
Task #:	1	2	...	16	
Classes per task:	5	5		5	20
Images in train split:	500	500	...	500	600
Images in test split:	100	100		100	

Table 4. The proposed 16-task split of Mini-ImageNet and CIFAR100 datasets for incremental few-shot learning.

methods mainly based on average accuracy as done.

References

[1] Q. Liu, O. Majumder, A. Achille, A. Ravichandran, R. Bhotika, S. Soatto, Incremental few-shot meta-learning via indirect discriminant alignment, in: European Conference on Computer Vision, Springer, 2020, pp. 685–701. 1, 2

[2] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, C. H. Lampert, icarl: Incremental classifier and representation learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2001–2010. 2

[3] J. Bang, H. Kim, Y. Yoo, J.-W. Ha, J. Choi, Rainbow memory: Continual learning with a memory of diverse samples, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 8218–8227. 2

1-shot/5-way ***4-task*** setting												
Learner:	ProtoNets											
Dataset:	Mini-ImageNet				CIFAR100				CUB			
Backbone:	4-Conv											
	Upper bound: 53.2				Upper bound: 55.4				Upper bound: 61.1			
Sessions:	1	2	3	4	1	2	3	4	1	2	3	4
FT	43.8	44.1	42.4	37.9	44.6	45.1	48.0	45.5	45.1	54.6	54.9	58.8
IDA	43.8	48.3	47.2	42.3	44.6	48.0	51.3	47.6	45.1	**54.7**	54.9	58.7
EIML	43.8	48.8	49.4	47.5	44.6	48.0	52.0	51.7	45.1	53.4	55.0	58.9
ERD	43.8	**51.1**	**52.3**	**53.0**	44.6	**49.5**	**53.6**	**55.1**	45.1	53.9	**58.3**	**60.8**
Backbone:	ResNet-12											
	Upper bound: 59.9				Upper bound: 61.8				Upper bound: 74.8			
Sessions:	1	2	3	4	1	2	3	4	1	2	3	4
FT	45.7	45.9	42.1	37.7	47.0	45.0	51.0	44.6	53.4	64.0	63.7	66.8
IDA	45.7	53.0	53.7	47.6	47.0	53.6	59.2	54.8	53.4	64.4	68.8	73.3
EIML	45.7	53.2	56.5	55.8	47.0	53.3	58.3	57.8	53.4	62.8	69.1	73.3
ERD	45.7	**55.2**	**58.2**	**59.3**	47.0	**55.6**	**61.3**	**61.4**	53.4	**66.1**	**72.4**	**74.1**

Table 5. Meta-test accuracy by training session in the 4-task setting. We evaluate *1-shot* and *5-way* few-shot recognition on three datasets using two different backbones.

5-shot/5-way ***4-task*** setting												
Learner:	ProtoNets											
Dataset:	Mini-ImageNet				CIFAR100				CUB			
Backbone:	4-Conv											
	Upper bound: 75.1				Upper bound: 76.5				Upper bound: 82.5			
Sessions:	1	2	3	4	1	2	3	4	1	2	3	4
FT	63.4	64.1	65.2	62.1	67.0	68.2	71.2	67.5	69.4	73.4	74.5	76.4
IDA	63.4	68.5	68.1	66.0	67.0	70.3	72.8	69.6	69.4	75.4	76.0	78.6
EIML	63.4	69.1	70.3	70.2	67.0	70.7	73.6	72.7	69.4	75.2	78.2	79.0
ERD	63.4	**69.4**	**71.4**	**72.2**	67.0	**71.2**	**74.4**	**73.9**	69.4	**75.9**	**78.7**	**80.4**
Backbone:	ResNet-12											
	Upper bound: 81.9				Upper bound: 81.0				Upper bound: 91.2			
Sessions:	1	2	3	4	1	2	3	4	1	2	3	4
FT	66.2	65.6	62.7	61.4	69.1	68.5	70.4	66.4	76.3	80.7	80.3	84.2
IDA	66.2	73.1	75.5	74.5	69.1	75.5	77.9	78.8	76.3	81.8	84.4	86.7
EIML	66.2	74.6	77.5	78.3	69.1	76.8	78.6	80.3	76.3	83.2	86.1	88.3
ERD	66.2	**74.7**	**77.7**	**80.0**	69.1	**77.2**	**79.4**	**80.8**	76.3	**83.4**	**86.6**	**89.6**

Table 6. Meta-test accuracy by training session in the 4-task setting. We evaluate *5-shot* and *5-way* few-shot recognition on three datasets using two different backbones.

Learner:	Relation Networks											
Datasets:	Mini-ImageNet				CIFAR100				CUB			
Backbone:	4-Conv											
1-shot/5-way ***16-task*** setting												
	Upper bound: 52.0				Upper bound: 59.2				Upper bound: 51.6			
Sessions:	2	4	8	16	2	4	8	16	2	4	8	16
FT	24.4	28.5	25.5	28.7	31.1	30.0	35.2	26.8	37.1	38.1	37.0	34.0
ERD	**27.7**	**29.9**	**34.5**	**30.1**	**35.6**	**39.3**	**45.7**	**35.9**	**37.3**	**42.9**	**47.9**	**42.5**
1-shot/5-way ***4-task*** setting												
	Upper bound: 52.0				Upper bound: 59.2				Upper bound: 51.6			
Sessions:	1	2	3	4	1	2	3	4	1	2	3	4
FT	41.70	41.65	38.51	33.33	42.9	45.3	45.7	42.3	45.7	47.9	48.5	50.3
ERD	41.70	**45.84**	**48.35**	**49.73**	42.9	**48.2**	**51.2**	**51.4**	45.7	**48.5**	**49.4**	**51.4**

Table 7. Meta-test accuracy by training sessions on the 4-task and 16-task settings. We evaluate 1-shot/5-way few-shot recognition on Mini-ImageNet, CIFAR-100 and CUB.