A Appendix

A.1 Details About Data Augmentation Techniques

In this section, we provide more details about the different data augmentation techniques we use in this work. We employ the following pool of data augmentation techniques:

CutMix: [25] introduced the CutMix augmentation strategy where patches are cut and pasted among training images, and the ground truth labels are also mixed proportionally to the area of the patches.

MixUp: [26] proposed mixup, a simple learning principle to alleviate memorization and sensitivity to adversarial examples. Mixup trains a neural network on convex combinations of pairs of examples and their labels. By doing so, mixup regularizes the neural network to favor simple linear behavior in between training examples.

Self-Mix: [20] introduced the self-mix augmentation strategy in which a patch of an image is substituted into other values in the same image to improve the generalization ability of few-shot image classification models.

In addition, we use some standard and simple data augmentation techniques:

Rotation: augments the data by rotating the images.

Horizontal Flip: augments the data by horizontally flipping images.

Random Erase: augments the data by randomly erasing patches from the image.

Finally, we also experimented with the following data augmentation techniques:

Combining Labels: augments the data by combining two different labels into a single class. For instance, we combine may combine the "dog" and "cat" labels to create a new "dog or cat" class.

Feature Mixup: similar to the "Mixup" augmentation technique we describe above, however we perform the mixup strategy on the feature representation for the image.

Drop Channel: augments the data by dropping color channels in the image.

Solarize: inverts all pixels above a threshold value of magnitude.

A.2 Detailed algorithm for Meta-MaxUp

Detailed algorithm for our proposed Meta-MaxUp. Algorithm 1 contains a more thorough description of this pipeline in practice (adapted from the standard meta-learning algorithm in [8]).

Algorithm 1 Meta-MaxUp

Require: Base model, F_θ, fine-tuning algorithm, $\mathcal{A}$, learning rate, γ set of augmentations $\mathcal{S}$, and distribution over tasks, $p(\mathcal{T})$.
Initialize θ, the weights of F;
while not done **do**
  Sample batch of tasks, $\{\mathcal{T}_i\}_{i=1}^{n}$, where $\mathcal{T}_i \sim p(\mathcal{T})$ and $\mathcal{T}_i = (\mathcal{T}_i^s, \mathcal{T}_i^q)$.
  for $i = 1, ..., n$ **do**
    Sample m augmentations, $\{M_j\}_{j=1}^{m}$, from $\mathcal{S}$.
    Compute $k = \arg\max_j \mathcal{L}(F_{\theta_j}, M_j(\mathcal{T}_i^q))$, where $\theta_j = \mathcal{A}(\theta, M_j(\mathcal{T}_i^s))$.
    Compute gradient $g_i = \nabla_\theta \mathcal{L}(F_{\theta_k}, M_k(\mathcal{T}_i^q))$.
  end for
  Update base model parameters: $\theta \leftarrow \theta - \frac{\gamma}{n} \sum_i g_i$.
end while

Table 4: Few-shot classification accuracy (%) on the CIFAR-FS dataset for all data augmentations. Confidence intervals have radius equal to one standard error. "CNN-4" denotes a 4-layer convolutional network with 96, 192, 384, and 512 filters in each layer [2]. Best performance in each category is bolded.

		CNN-4		ResNet-12	
Mode	Level	1-shot	5-shot	1-shot	5-shot
Baseline	-	67.56 ± 0.35	82.39 ± 0.26	73.01 ± 0.37	84.29 ± 0.24
Random Erase	Support	67.71 ± 0.36	82.25 ± 0.26	72.30 ± 0.37	84.50 ± 0.25
Self-Mix	Support	**69.61 ± 0.35**	**83.43 ± 0.25**	71.96 ± 0.36	**84.84 ± 0.25**
CutMix	Support	69.05 ± 0.36	83.12 ± 0.26	**72.60 ± 0.37**	84.70 ± 0.25
MixUp	Support	68.64 ± 0.37	82.72 ± 0.27	71.86 ± 0.37	84.11 ± 0.25
Feature Mixup	Support	67.88 ± 0.35	82.40 ± 0.25	71.21 ± 0.37	83.38 ± 0.25
Rotation	Support	68.65 ± 0.35	82.86 ± 0.25	71.13 ± 0.37	83.84 ± 0.25
Combining labels	Support	68.27 ± 0.36	82.53 ± 0.26	71.00 ± 0.38	83.12 ± 0.25
Drop Channel	Support	68.21 ± 0.35	82.76 ± 0.25	69.65 ± 0.73	83.15 ± 0.25
Solarize	Support	68.65 ± 0.35	82.68 ± 0.26	70.88 ± 0.37	83.45 ± 0.25
Random Erase	Query	69.73 ± 0.34	84.04 ± 0.25	73.05 ± 0.36	85.67 ± 0.25
Self-Mix	Query	69.61 ± 0.35	83.43 ± 0.25	71.96 ± 0.36	84.84 ± 0.25
CutMix	Query	**70.54 ± 0.33**	**84.69 ± 0.24**	**75.97 ± 0.34**	**87.28 ± 0.23**
MixUp	Query	67.70 ± 0.34	83.13 ± 0.25	72.93 ± 0.35	86.13 ± 0.24
Feature Mixup	Query	70.16 ± 0.35	83.80 ± 0.28	73.38 ± 0.35	85.87 ± 0.23
Rotation	Query	68.17 ± 0.35	83.01 ± 0.25	72.02 ± 0.36	84.42 ± 0.25
Combining labels	Query	66.01 ± 0.34	81.99 ± 0.26	69.77 ± 0.37	82.99 ± 0.26
Drop Channel	Query	68.34 ± 0.35	83.25 ± 0.25	69.60 ± 0.37	83.01 ± 0.26
Solarize	Query	67.51 ± 0.35	82.65 ± 0.25	72.45 ± 0.36	84.97 ± 0.24
MixUp	Task	67.21 ± 0.35	82.72 ± 0.26	72.05 ± 0.37	85.27 ± 0.25
Large Rotation	Task	**68.96 ± 0.35**	**83.65 ± 0.25**	**73.79 ± 0.36**	**85.81 ± 0.24**
CutMix	Task	68.78 ± 0.36	82.99 ± 0.50	72.72 ± 0.37	84.62 ± 0.25
Combining labels	Task	68.08 ± 0.35	82.33 ± 0.26	69.64 ± 0.37	83.79 ± 0.26
Random Erase	Task	68.39 ± 0.36	83.26 ± 0.25	71.09 ± 0.37	84.49 ± 0.25
Drop Channel	Task	67.54 ± 0.36	81.97 ± 0.25	70.24 ± 0.37	83.52 ± 0.26
Horizontal Flip	Shot	**68.13 ± 0.35**	82.95 ± 0.25	**73.25 ± 0.36**	**85.06 ± 0.25**
Random Crop	Shot	67.33 ± 0.36	**83.04 ± 0.25**	70.56 ± 0.37	83.87 ± 0.25
Random Rotation	Shot	67.57 ± 0.35	83.00 ± 0.25	70.32 ± 0.37	83.75 ± 0.25

A.3 Results for All Data Augmentation Techniques

Table 4 contains results for a bunch of data augmentation in different modes, namely Support, Query, Task and Shot.

A.4 Experimental Details

The mini-ImageNet dataset consists of 100 randomly chosen classes from ILSVRC-2012 [19]. These classes are randomly splited into training classes, 16 validation classes, and 20 classes for testing. CIFAR-FS consists of all 100 classes from CIFAR-100, and splited into training classes, 16 validation classes, and 20 classes for testing as well.

For MetOptNet, we use the same training procedure as [14] including SGD with Nesterov momentum of 0.9 and weight decay coefficient 0.0005. The model was meta-trained for 60 epochs, with an initial learning rate 0.1, then changed to 0.006, 0.0012, and 0.00024 at epochs 20, 40 and 50, respectively. In each epoch, we train on 8000 episodes and use mini-batches of size 8. Following [14], we use a larger shot number (15) to train mini-ImageNet for both 1-shot and 5-shot classification. For MCT, we use the same optimizer but with batch size 1 and maximum iterations 50000. Following [13], we enlarge the training classification ways to 15 for a 5-way testing. We use instance-wise metric for all inductive learning.

A.5 Results for Using Combination of Data Augmentation

Table 5 collects results for combination effective data augmentations with the most effective method, namely CutMix in Query mode.

Table 5: Few-shot classification accuracy (%) on the CIFAR-FS dataset with combinations of augmentations and query CutMix. "S","Q","T" denote "Support", "Query", and "Task" modes, respectively. While adding augmentations can help, it can also hurt, so additional augmentations must be chosen carefully.

	CNN-4		ResNet-12	
Mode	1-shot	5-shot	1-shot	5-shot
CutMix	70.54 ± 0.33	84.69 ± 0.24	75.97 ± 0.34	87.28 ± 0.23
+ CutMix (S)	70.00 ± 0.35	84.41 ± 0.25	75.87 ± 0.35	87.05 ± 0.25
+ Random Erase (S)	70.12 ± 0.35	84.48 ± 0.25	75.84 ± 0.34	87.19 ± 0.24
+ Random Erase (Q)	69.68 ± 0.34	84.36 ± 0.24	75.08 ± 0.35	87.14 ± 0.23
+ Self-Mix (S)	70.65 ± 0.34	84.68 ± 0.25	$\mathbf{76.27 \pm 0.34}$	87.52 ± 0.24
+ Self-Mix (Q)	69.94 ± 0.34	84.38 ± 0.24	76.04 ± 0.34	87.45 ± 0.24
+ MixUp (T)	70.33 ± 0.35	84.57 ± 0.25	75.97 ± 0.34	86.66 ± 0.24
+ Rotation (T)	70.35 ± 0.34	84.73 ± 0.24	75.74 ± 0.34	$\mathbf{87.68 \pm 0.24}$
+ Horizontal Flip (Shot)	$\mathbf{70.90 \pm 0.33}$	$\mathbf{84.87 \pm 0.24}$	76.23 ± 0.34	87.36 ± 0.24

A.6 Results for Backbones Other Than ResNet-12

Table 6 shows the results for backbones other than ResNet-12. Training with proposed data augmentation and Meta-MaxUp again outforms the baselines by a large margin (1%-3%).

Table 6: Few-shot classification accuracy (%) on CIFAR-FS and mini-ImageNet. "+ DA" denotes training with CutMix (Q) + Rotation (T), and "+ MM" denotes training with Meta-MaxUp. "64-64-64-64" denotes the 4-layer CNN backbone from [21].

		CIFAR-FS		mini-ImageNet	
Method	Backbone	1-shot	5-shot	1-shot	5-shot
R2-D2	CNN-4	67.56 ± 0.35	82.39 ± 0.26	56.15 ± 0.31	72.46 ± 0.26
+ DA	CNN-4	70.54 ± 0.33	84.69 ± 0.24	57.60 ± 0.32	74.69 ± 0.25
+ MM	CNN-4	$\mathbf{71.10 \pm 0.34}$	$\mathbf{85.50 \pm 0.24}$	$\mathbf{58.18 \pm 0.32}$	$\mathbf{75.35 \pm 0.25}$
ProtoNet	64-64-64-64	60.91 ± 0.35	79.73 ± 0.27	47.97 ± 0.32	70.13 ± 0.27
+ DA	64-64-64-64	62.21 ± 0.36	80.70 ± 0.27	$\mathbf{50.38 \pm 0.32}$	$\mathbf{71.44 \pm 0.26}$
+ MM	64-64-64-64	$\mathbf{63.01 \pm 0.36}$	$\mathbf{80.85 \pm 0.25}$	50.06 ± 0.32	71.13 ± 0.26

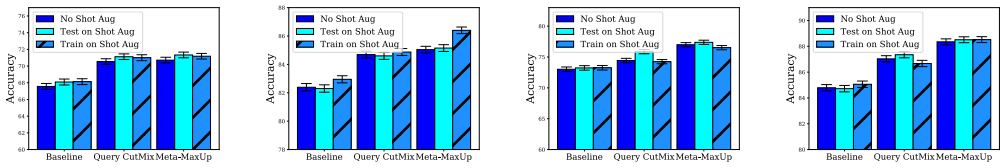
A.7 Augmentation Pool for Meta-MaxUp

For all the benchmark results of Meta-MaxUp training, we use a medium-size data augmentation pool with $m = 4$, including CutMix (Q), Random Erase (Q), Self-Mix (S), Rotation (T), CutMix (Q) + Rotation (T), and Random Erase (Q) + Rotation (T). For the large-size pool, we add more techniques and combinations of the mentioned techniques into the pool, including Random Erase (Q) + Random Erase (S), CutMix (Q) + Random Erase (S), CutMix (Q) + Random Erase (Q), and CutMix (Q) + Self-Mix (S).

Table 7 shows the results for various values of m and different data augmentation pool sizes. Rows with $m = 1$ denote experiments where we do not maximize loss in the inner loop and thus simply apply randomly sampled data augmentation for each task.

Table 7: Few-shot classification accuracy (%) on the CIFAR-FS dataset for Meta-MaxUp over different sizes of augmentation pools and numbers of samples. As m and the pool size increase, so does performance. Meta-MaxUp is able to pick effective augmentations from a large pool.

		CNN-4		ResNet-12	
Pool	m	1-shot	5-shot	1-shot	5-shot
Baseline	-	67.56 ± 0.36	82.39 ± 0.26	73.01 ± 0.37	84.29 ± 0.24
CutMix	1	70.54 ± 0.34	84.69 ± 0.24	75.97 ± 0.34	87.28 ± 0.23
Single	1	70.76 ± 0.35	84.70 ± 0.25	75.71 ± 0.35	87.44 ± 0.43
Medium	1	70.50 ± 0.34	84.59 ± 0.24	75.60 ± 0.34	87.35 ± 0.23
Large	1	70.84 ± 0.34	85.04 ± 0.24	75.44 ± 0.34	87.47 ± 0.23
CutMix	2	70.56 ± 0.34	84.78 ± 0.24	74.93 ± 0.36	87.14 ± 0.24
Single	2	70.86 ± 0.34	85.06 ± 0.25	75.81 ± 0.34	87.33 ± 0.23
Medium	2	70.75 ± 0.34	85.02 ± 0.24	76.49 ± 0.33	88.20 ± 0.22
Large	2	70.63 ± 0.34	85.07 ± 0.24	76.59 ± 0.34	88.11 ± 0.23
CutMix	4	70.48 ± 0.34	84.76 ± 0.24	75.08 ± 0.23	87.60 ± 0.24
Single	4	**71.10 ± 0.34**	**85.50 ± 0.24**	76.82 ± 0.24	88.14 ± 0.23
Medium	4	70.58 ± 0.34	85.32 ± 0.24	76.30 ± 0.24	88.29 ± 0.22
Large	4	70.71 ± 0.34	85.04 ± 0.23	**76.99 ± 0.24**	**88.35 ± 0.22**



(a) (b) (c) (d)

Figure 3: Performance for shot augmentation for different backbone works and training strategies on CIFAR-FS. (a) 1-shot classification for CNN-4 (b) 5-shot classification for CNN-4 (c) 1-shot classification for ResNet-12 (d) 5-shot classification for ResNet-12

A.8 Bar Plots for Shot Augmentation

A.9 Compare with Existing Data Augmentation Methods for Meta-learning

Table 8 compares our proposed Meta-MaxUp with Large Rotation in task mode [15].

Table 8: Few-shot classification accuracy (%) on CIFAR-FS and mini-ImageNet with ResNet-12 backbone. "M-SVM" denotes MetaOptNet with the SVM head. "+ens" denotes testing with ensemble methods as in [15]. "LargeRot" denotes task-level augmentation by Large Rotations as described in [15].

	CIFAR-FS		mini-ImageNet	
Method	1-shot	5-shot	1-shot	5-shot
M-SVM + LargeRot	72.95 ± 0.24	85.91 ± 0.18	62.12 ± 0.22	78.90 ± 0.17
M-SVM + LargeRot + ens	75.85 ± 0.24	87.73 ± 0.17	64.56 ± 0.22	81.35 ± 0.16
M-SVM + MM (ours)	75.67 ± 0.34	88.37 ± 0.23	65.02 ± 0.32	82.42 ± 0.23
M-SVM + MM + ens (ours)	76.38 ± 0.33	89.16 ± 0.22	66.42 ± 0.32	83.69 ± 0.21
M-SVM + LargeRot + ens + val	**76.75 ± 0.23**	88.38 ± 0.17	65.38 ± 0.23	82.13 ± 0.16
M-SVM + MM + ens + val (ours)	76.38 ± 0.34	**89.25 ± 0.21**	**67.37 ± 0.32**	**84.57 ± 0.21**