A Additional experiment details

We refer the reader to the VTAB[12] and Meta-Dataset[13] codebases for complete experimental details. Instructions on how to evaluate new approaches on VTAB+MD can be found at https://github.com/google-research/meta-dataset/blob/main/VTAB-plus-MD.md.

Experiments presented in this work are ran in two main computing infrastructure: TPU-v3 (all BIT experiments) and Nvidia V100 (rest).

For Prototypical networks, ProtoMAML and MD-Transfer, model and hyperparameter selection is based on the average query accuracy over episodes sampled from all of MD-v2's validation classes. For each approach we perform a hyperparameter search using Triantafillou et al. [60]'s search space (Tables 1, 2, and 3, presented alongside the best values found), for a total of 99 runs for each approach.

We re-train CrossTransformers on episodes sampled from all ImageNet classes, with 50% of the episodes converted to SimCLR episodes — this corresponds to the CTX+SimCLR Eps setting in Doersch et al. [14]. We use the recommended hyperparameters and perform a light sweep over learning rates in {0.01, 0.001, 0.0006, 0.0001} and found Doersch et al. [14]'s recommended 0.0006 learning rate to be optimal in our case as well. Model selection is performed using MD-v2 validation episodes — this is a slight departure from CrossTransformers' ImageNet-only prototol that is made necessary by the fact that all ImageNet classes participate in training episodes in MD-v2.

Since pre-trained SUR backbones were already made available by the authors,[14] we re-used all of them with two exceptions: (1) we re-trained the ImageNet backbone on all ImageNet classes using the provided training script (because the original backbone was trained on Meta-Dataset's ImageNet training classes), and (2) we ignored the VGG Flowers backbone (because the dataset is included as one of VTAB-v2's downstream tasks). We ran Dvornik et al. [16]'s inference code as-is for evaluation.

All Big Transfer models are pre-trained as described in [31]. The pre-processing at training time is at 224 resolution, using random horizontal flipping and inception crop [57]. In all of our experiments, during transfer we only resize images to the desired resolution (126 or 224) at both fine-tuning and evaluation time. While higher resolution and further data augmentation further improves performance, we remove this additional confounding factor.

Hyperparameter	Search space	Best
Backbone	{ResNet-18, 4-layer convnet}	ResNet-18
Resolution	{84, 126}	126
Outer-loop LR	log-uniform(1e-6, 1e-2)	0.0004
Outer-loop LR decay freq.	{100, 500, 1k, 2.5k, 5k, 10k}	1k
Outer-loop LR decay rate	uniform(0.5, 1.0)	0.6478
Inner-loop LR	log-uniform(5e-3, 5e-1)	0.0054
Inner-loop steps	{1, 6, 10}	10
Additional inner-loop steps (evaluation)	{0, 5}	0

Table 1: ProtoMAML hyperparameter search space.

B Detailed figures and accuracy tables

We show a detailed breakdown of VTAB-V2 accuracies (Figure 7) for investigated approaches. We also provide detailed accuracy tables (Tables 4 through 9) for all plots displayed in the main text. For MD-v2 we show 95% confidence intervals computed over 60 episodes for BiT learners and 600 episodes for all other approaches.

[12] https://github.com/google-research/task_adaptation

[13] https://github.com/google-research/meta-dataset

[14] https://github.com/dvornikita/SUR

Hyperparameter	Search space	Best
Backbone	{ResNet-18, 4-layer convnet}	ResNet-18
Resolution	{84, 126}	126
Training LR	log-uniform(1e-6, 1e-2)	3.4293725734843445e-06
Fine-tuning LR	{1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 2e-1}	1e-2
Fine-tuning steps	{50, 75, 100, 125, 150, 175, 200}	100
Fine-tune with Adam?	{True, False}	True
Cosine classifier head?	{True, False}	True
Cosine logits multiplier	{1, 2, 10, 100}	10
Weight-normalize the classifier head?	{True, False}	True
Fine-tune all layers?	{True, False}	True

Table 2: MD-Transfer hyperparameter search space.

Hyper	Search space	Best
Backbone	{ResNet-18, 4-layer convnet}	ResNet-18
Resolution	{84, 126}	126
LR	log-uniform(1e-6, 1e-2)	0.0003
LR decay freq.	{100, 500, 1k, 2.5k, 5k, 10k}	500
LR decay rate	uniform(0.5, 1.0)	0.8857

Table 3: Prototypical Networks hyperparameter search space.

C Bridging the Performance Gap Between MD-Transfer Baseline and ProtoMAML

Given the stark differences between ProtoMAML and MD-Transfer on VTAB-v2, we ran a few additional experiments in order to better explain these discrepancies. We swapped their evaluation hyperparameters, meaning that we fine-tuned MD-Transfer for 10 steps using a learning rate of 0.0054 without using a cosine classifier (**MD-Transfer (ProtoMAML hypers)**) and that we ran ProtoMAML's inner-loop for 100 steps using a learning rate of 1×10^{-2} with a linear classification head (**ProtoMAML (MD-Transfer hypers)**). Note that this does not completely bridge the hyperparameter gap between the two approaches, but it does bring them closer to each other. The remaining differences are that (1) the validation procedure used for early stopping is different, and (2) ProtoMAML initializes the output layer with class prototypes, whereas the output layer weights in MD-Transfer are sampled from a normal distribution. Additionally, to isolate the effect of cosine-classification, we run MD-Transfer with a linear classification head while keeping the learning rate and number of training steps the same (**MD-Transfer (linear head)**).

Figure 8 shows that ProtoMAML gets better results on MD-v2 with MD-Transfer hyperparameters (more fine-tuning steps with a smaller learning rate), with apparent gains on Quickdraw and Traffic Signs. ProtoMAML's prototypical initialization seems to yield better performance for "in-domain" datasets (i.e. datasets participating to the training split of classes), however we observe diminishing returns for test-only datasets like Traffic Sign.

Disabling cosine classification (**MD-Transfer (linear head)**) seems to harm fine-tuning performance greatly on all datasets except QuickDraw. Traffic Signs in particuar benefits greatly from a cosine classification head, as evidenced by the 10% drop in performance observed when switching to a linear classification head. On VTAB, again, MD-Transfer hyperparameters help improve ProtoMAML performance, hinting at the fact that the hyperparameter selection procedure used for ProtoMAML is sub-optimal.

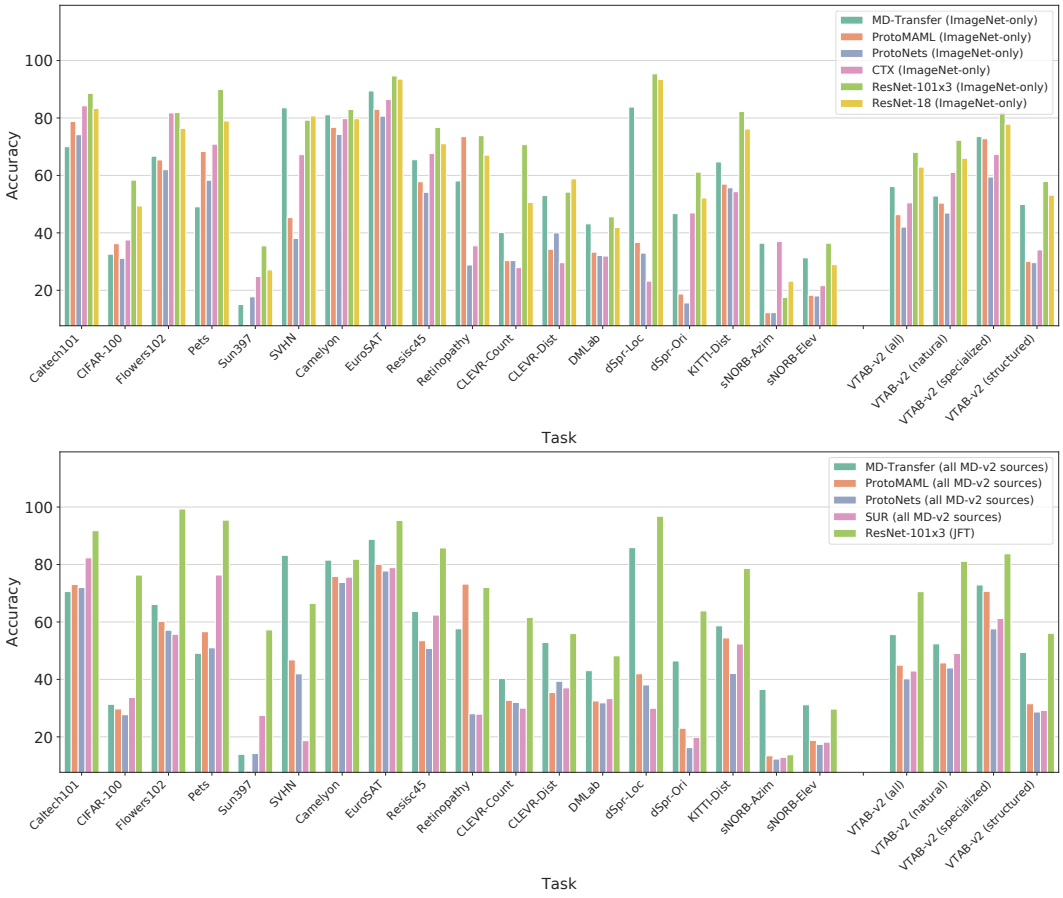

Figure 7: VTAB-v2 accuracies, broken down by downstream task, for approaches trained only on ImageNet (*top*) or larger-scale datasets (*bottom*).

D Larger-scale SUR experiments

In this section we investigate increasing the capacity (ResNet-50) and input resolution (224×224) of SUR backbones. We re-train backbones for all seven of MD-v2's training sources of data using BiT's upstream training hyperparameters and adjusting the number of training steps as needed to ensure convergence. We trained two backbone variants: one with a regular linear classification head, and one with a temperature-adjusted cosine classifier head. Backbones were trained for:

- **ImageNet**: 90 epochs
- **Quickdraw**: 4 epochs
- **Birds, Omniglot, Fungi**: 900 epochs
- **Textures**: 1350 epochs
- **Aircraft**: 4500 epochs

The LR schedule is adjusted proportionally to the number of epochs. For simplicity we select the final backbone checkpoints rather than selecting based on an episodic loss.

Figure 9 shows an appreciable 5% improvement on VTAB-v2, most of which is driven by an improvement on specialized tasks. On the other hand, the aggregate performance gain on MD-v2 is negligible. While performance on MSCOCO, Fungi, Birds, and Textures is increased significantly, the larger input resolution and backbone capacity has a negligible or detrimental effect on QuickDraw, Omniglot, and Aircraft. We hypothesize that the drop in Aircraft performance is due to the large batch size used by BiT and a suboptimal model selection strategy.

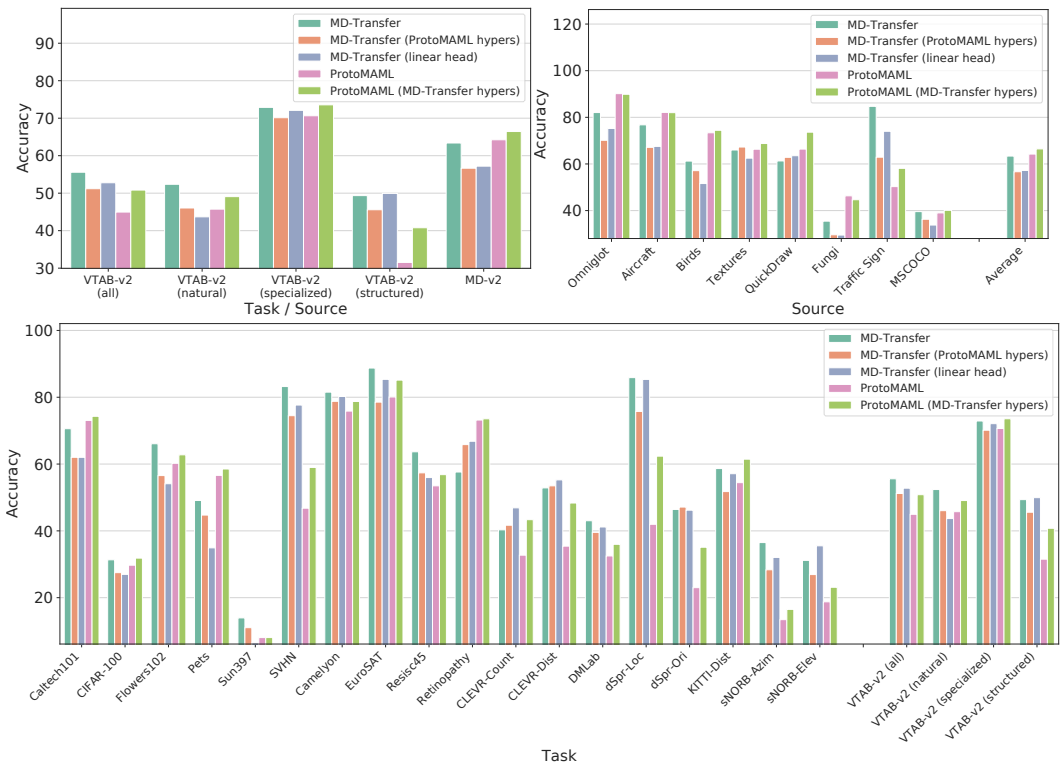

Figure 8: Ablation study for different hyper parameters found by ProtoMAML and MD-Transfer, broken down by downstream task. All backbones are trained the all MD-V2 training data.

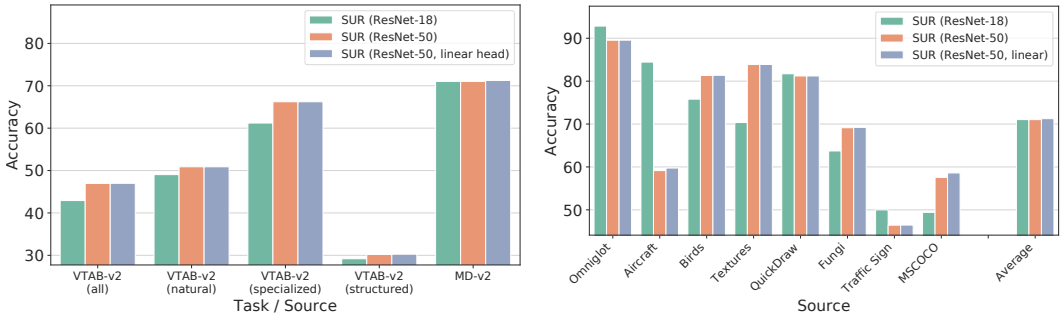

Figure 9: Ablation study for different hyper parameters found by ProtoMAML and MD-Transfer, broken down by downstream task, for Meta Dataset-v2 (*top*) and VTAB (*bottom*). All backbones are trained the all MD-V2 training data.

Overall these results are encouraging, but a more thorough investigation is needed before we can draw definitive conclusions.

Data source	MD-Transfer	ProtoMAML	ProtoNets	CTX	BiT-ResNet-101x3	BiT-ResNet-18
Omniglot	$80.92 \pm 1.20\%$	$68.35 \pm 1.28\%$	$65.47 \pm 1.35\%$	$84.55 \pm 0.94\%$	$72.35 \pm 4.70\%$	$71.87 \pm 4.38\%$
Aircraft	$75.45 \pm 1.20\%$	$58.18 \pm 0.96\%$	$54.25 \pm 1.03\%$	$85.31 \pm 0.83\%$	$78.34 \pm 3.57\%$	$70.23 \pm 3.78\%$
Birds	$61.23 \pm 1.30\%$	$69.69 \pm 0.98\%$	$64.78 \pm 0.98\%$	$72.92 \pm 1.07\%$	$91.02 \pm 1.49\%$	$81.65 \pm 2.26\%$
DTD	$66.66 \pm 1.01\%$	$68.71 \pm 0.83\%$	$64.91 \pm 0.76\%$	$77.29 \pm 0.71\%$	$87.06 \pm 2.61\%$	$78.62 \pm 2.86\%$
QuickDraw	$61.12 \pm 1.06\%$	$55.52 \pm 1.02\%$	$53.26 \pm 1.02\%$	$73.29 \pm 0.78\%$	$65.08 \pm 4.13\%$	$64.81 \pm 3.71\%$
Fungi	$35.39 \pm 1.08\%$	$38.88 \pm 1.05\%$	$36.37 \pm 1.08\%$	$47.95 \pm 1.19\%$	$60.68 \pm 4.43\%$	$49.81 \pm 4.28\%$
Traffic Sign	$85.31 \pm 0.95\%$	$53.83 \pm 1.05\%$	$50.27 \pm 1.05\%$	$80.12 \pm 0.97\%$	$76.23 \pm 4.68\%$	$69.53 \pm 4.55\%$
MSCOCO	$39.66 \pm 1.05\%$	$43.32 \pm 1.12\%$	$41.08 \pm 0.99\%$	$51.39 \pm 1.06\%$	$69.74 \pm 2.69\%$	$57.84 \pm 3.03\%$
Caltech101	70.00 %	78.81 %	74.18 %	84.24 %	88.59 %	83.32 %
CIFAR100	32.57 %	36.22 %	31.13 %	37.51 %	58.35 %	49.37 %
Flowers102	66.69 %	65.39 %	61.99 %	81.75 %	81.88 %	76.38 %
Pets	49.06 %	68.33 %	58.33 %	70.88 %	89.97 %	78.95 %
Sun397	15.05 %	8.05 %	17.73 %	24.79 %	35.47 %	27.09 %
SVHN	83.54 %	45.31 %	38.06 %	67.22 %	79.23 %	80.71 %
EuroSAT	89.41 %	83.02 %	80.63 %	86.43 %	94.64 %	93.53 %
Resics45	65.46 %	57.79 %	54.11 %	67.65 %	76.71 %	71.03 %
Patch Camelyon	81.11 %	76.75 %	74.26 %	79.77 %	82.97 %	79.73 %
Retinopathy	58.07 %	73.51 %	28.82 %	35.48 %	73.85 %	67.06 %
CLEVR-count	40.09 %	30.32 %	30.33 %	27.89 %	70.73 %	50.59 %
CLEVR-dist	52.97 %	34.29 %	39.99 %	29.61 %	54.19 %	58.79 %
dSprites-loc	83.81 %	36.68 %	32.95 %	23.19 %	95.38 %	93.39 %
dSprites-ori	46.70 %	18.69 %	15.60 %	46.92 %	61.13 %	52.15 %
SmallNORB-azi	36.40 %	12.20 %	12.21 %	37.02 %	17.50 %	23.17 %
SmallNORB-elev	31.29 %	18.26 %	18.02 %	21.62 %	36.40 %	28.92 %
DMLab	43.14 %	33.28 %	32.12 %	31.92 %	45.58 %	41.86 %
KITTI-dist	64.70 %	56.96 %	55.70 %	54.34 %	82.24 %	76.15 %
MD-v2	63.22 %	57.06 %	53.80 %	71.60 %	75.06 %	68.04 %
VTAB (all)	56.11 %	46.33 %	42.01 %	50.46 %	68.04 %	62.90 %
VTAB (natural)	52.82 %	50.35 %	46.90 %	61.07 %	72.25 %	65.97 %
VTAB (specialized)	73.51 %	72.77 %	59.45 %	67.33 %	82.04 %	77.84 %
VTAB (structured)	49.89 %	30.08 %	29.62 %	34.06 %	57.89 %	53.13 %

Table 4: VTAB+MD accuracies for approaches trained only on ImageNet.

Data source	MD-Transfer	ProtoMAML	ProtoNets	SUR	BiT-ResNet-101x3 (JFT)
Omniglot	$82.04 \pm 1.27\%$	$90.15 \pm 0.65\%$	$85.29 \pm 0.89\%$	$92.84 \pm 0.52\%$	$76.45 \pm 4.04\%$
Aircraft	$76.77 \pm 1.16\%$	$82.10 \pm 0.60\%$	$74.34 \pm 0.81\%$	$84.44 \pm 0.58\%$	$93.30 \pm 1.44\%$
Birds	$61.23 \pm 1.29\%$	$73.36 \pm 0.92\%$	$68.00 \pm 1.01\%$	$75.80 \pm 0.96\%$	$97.06 \pm 0.53\%$
DTD	$65.98 \pm 1.07\%$	$66.32 \pm 0.76\%$	$65.26 \pm 0.69\%$	$70.35 \pm 0.72\%$	$88.96 \pm 2.14\%$
QuickDraw	$61.29 \pm 1.06\%$	$66.37 \pm 0.95\%$	$60.57 \pm 1.00\%$	$81.71 \pm 0.57\%$	$71.27 \pm 3.77\%$
Fungi	$35.47 \pm 1.05\%$	$46.32 \pm 1.11\%$	$39.84 \pm 1.10\%$	$63.72 \pm 1.08\%$	$62.59 \pm 4.29\%$
Traffic Sign	$84.71 \pm 0.94\%$	$50.28 \pm 1.05\%$	$49.79 \pm 1.07\%$	$49.99 \pm 1.08\%$	$69.13 \pm 5.34\%$
MSCOCO	$39.56 \pm 1.00\%$	$39.00 \pm 1.04\%$	$39.65 \pm 1.03\%$	$49.41 \pm 1.08\%$	$76.36 \pm 2.23\%$
Caltech101	70.58 %	73.06 %	71.98 %	82.33 %	91.78 %
CIFAR100	31.33 %	29.72 %	27.70 %	33.69 %	76.32 %
Flowers102	66.08 %	60.22 %	57.11 %	55.72 %	99.33 %
Pets	49.09 %	56.61 %	50.99 %	76.34 %	95.45 %
Sun397	13.94 %	8.05 %	14.19 %	27.49 %	57.24 %
SVHN	83.20 %	46.78 %	41.93 %	18.66 %	66.47 %
EuroSAT	88.74 %	80.07 %	77.74 %	78.91 %	95.33 %
Resics45	63.67 %	53.48 %	50.79 %	62.40 %	85.76 %
Patch Camelyon	81.53 %	75.85 %	73.75 %	75.60 %	81.81 %
Retinopathy	57.61 %	73.18 %	28.04 %	27.91 %	72.02 %
CLEVR-count	40.30 %	32.72 %	31.96 %	29.99 %	61.54 %
CLEVR-dist	52.86 %	35.43 %	39.35 %	37.06 %	55.96 %
dSprites-loc	85.87 %	41.96 %	38.07 %	29.96 %	96.80 %
dSprites-ori	46.41 %	23.00 %	16.25 %	19.84 %	63.84 %
SmallNORB-azi	36.49 %	13.42 %	12.27 %	12.86 %	13.78 %
SmallNORB-elev	31.16 %	18.76 %	17.38 %	18.15 %	29.68 %
DMLab	43.03 %	32.49 %	31.83 %	33.31 %	48.22 %
KITTI-dist	58.65 %	54.43 %	42.05 %	52.32 %	78.62 %
MD-v2	63.38 %	64.24 %	60.34 %	71.03 %	79.39 %
VTAB (all)	55.59 %	44.96 %	40.19 %	42.92 %	70.55 %
VTAB (natural)	52.37 %	45.74 %	43.98 %	49.04 %	81.10 %
VTAB (specialized)	72.89 %	70.65 %	57.58 %	61.20 %	83.73 %
VTAB (structured)	49.35 %	31.52 %	28.65 %	29.19 %	56.05 %

Table 5: VTAB+MD accuracies for approaches trained on more data (all of MD-v2's training sources, unless noted otherwise).

Data source	BiT-ResNet-18 (126 × 126)	BiT-ResNet-18 (224 × 224)	BiT-ResNet-50 (126 × 126)	BiT-ResNet-50 (224 × 224)	CTX
Omniglot	$71.87 \pm 4.38\%$	$72.73 \pm 4.64\%$	$68.56 \pm 4.68\%$	$68.03 \pm 4.86\%$	$84.55 \pm 0.94\%$
Aircraft	$70.23 \pm 3.78\%$	$73.61 \pm 3.80\%$	$74.09 \pm 3.64\%$	$77.42 \pm 3.55\%$	$85.31 \pm 0.83\%$
Birds	$81.65 \pm 2.26\%$	$87.22 \pm 1.88\%$	$86.82 \pm 1.57\%$	$90.82 \pm 1.46\%$	$72.92 \pm 1.07\%$
DTD	$78.62 \pm 2.86\%$	$82.62 \pm 2.70\%$	$82.35 \pm 2.56\%$	$84.97 \pm 2.53\%$	$77.29 \pm 0.71\%$
QuickDraw	$64.81 \pm 3.71\%$	$66.34 \pm 3.60\%$	$66.98 \pm 3.62\%$	$66.56 \pm 3.69\%$	$73.29 \pm 0.78\%$
Fungi	$49.81 \pm 4.28\%$	$53.93 \pm 4.44\%$	$54.63 \pm 4.20\%$	$59.37 \pm 4.25\%$	$47.95 \pm 1.19\%$
Traffic Sign	$69.53 \pm 4.55\%$	$75.39 \pm 4.34\%$	$71.09 \pm 4.66\%$	$73.52 \pm 4.69\%$	$80.12 \pm 0.97\%$
MSCOCO	$57.84 \pm 3.03\%$	$59.97 \pm 2.89\%$	$64.55 \pm 2.93\%$	$65.69 \pm 2.71\%$	$51.39 \pm 1.06\%$
Caltech101	83.32 %	84.59 %	85.69 %	87.22 %	84.24 %
CIFAR100	49.37 %	47.10 %	55.85 %	54.42 %	37.51 %
Flowers102	76.38 %	82.65 %	81.87 %	83.33 %	81.75 %
Pets	78.95 %	83.91 %	86.07 %	87.91 %	70.88 %
Sun397	27.09 %	29.11 %	31.62 %	33.29 %	24.79 %
SVHN	80.71 %	83.40 %	78.47 %	70.40 %	67.22 %
EuroSAT	93.53 %	93.82 %	94.14 %	94.44 %	86.43 %
Resics45	71.03 %	74.12 %	74.92 %	76.13 %	67.65 %
Patch Camelyon	79.73 %	80.67 %	81.55 %	83.06 %	79.77 %
Retinopathy	67.06 %	74.47 %	71.15 %	70.24 %	35.48 %
CLEVR-count	50.59 %	55.25 %	53.69 %	74.03 %	27.89 %
CLEVR-dist	58.79 %	58.69 %	54.59 %	51.55 %	29.61 %
dSprites-loc	93.39 %	98.59 %	92.53 %	82.72 %	23.19 %
dSprites-ori	52.15 %	46.46 %	51.40 %	55.11 %	46.92 %
SmallNORB-azi	23.17 %	20.71 %	20.10 %	17.79 %	37.02 %
SmallNORB-elev	28.92 %	21.75 %	26.95 %	32.07 %	21.62 %
DMLab	41.86 %	43.74 %	42.54 %	43.18 %	31.92 %
KITTI-dist	76.15 %	78.78 %	77.80 %	79.93 %	54.34 %
MD-v2	68.04 %	71.48 %	71.14 %	73.30 %	71.60 %
VTAB (all)	62.90 %	64.32 %	64.50 %	65.38 %	50.46 %
VTAB (natural)	65.97 %	68.46 %	69.93 %	69.43 %	61.07 %
VTAB (specialized)	77.84 %	80.77 %	80.44 %	80.97 %	67.33 %
VTAB (structured)	53.13 %	53.00 %	52.45 %	54.55 %	34.06 %

Table 6: VTAB+MD accuracies for BiT learners trained on various input resolutions and network capacities. CrossTransformers (CTX) accuracies are provided for context. All approaches are trained only on ImageNet.

Data source	BiT-ResNet-50 (GNWS)	BiT-ResNet-50 (BN)
Omniglot	$68.03 \pm 4.86\%$	$61.66 \pm 5.13\%$
Aircraft	$77.42 \pm 3.55\%$	$76.82 \pm 3.71\%$
Birds	$90.82 \pm 1.46\%$	$87.59 \pm 1.84\%$
DTD	$84.97 \pm 2.53\%$	$83.72 \pm 3.39\%$
QuickDraw	$66.56 \pm 3.69\%$	$63.83 \pm 4.03\%$
Fungi	$59.37 \pm 4.25\%$	$53.77 \pm 4.43\%$
Traffic Sign	$73.52 \pm 4.69\%$	$70.46 \pm 4.70\%$
MSCOCO	$65.69 \pm 2.71\%$	$61.50 \pm 2.73\%$
Caltech101	87.22 %	88.72 %
CIFAR100	54.42 %	53.78 %
Flowers102	83.33 %	85.45 %
Pets	87.91 %	88.24 %
Sun397	33.29 %	31.60 %
SVHN	70.40 %	85.57 %
EuroSAT	94.44 %	95.35 %
Resics45	76.13 %	79.02 %
Patch Camelyon	83.06 %	80.13 %
Retinopathy	70.24 %	73.13 %
CLEVR-count	74.03 %	43.10 %
CLEVR-dist	51.55 %	49.65 %
dSprites-loc	82.72 %	83.19 %
dSprites-ori	55.11 %	46.49 %
SmallNORB-azi	17.79 %	18.93 %
SmallNORB-elev	32.07 %	34.32 %
DMLab	43.18 %	44.67 %
KITTI-dist	79.93 %	76.97 %
MD-v2	73.30 %	69.92 %
VTAB (all)	65.38 %	64.35 %
VTAB (natural)	69.43 %	72.22 %
VTAB (specialized)	80.97 %	81.91 %
VTAB (structured)	54.55 %	49.67 %

Table 7: VTAB+MD accuracies for BiT learners trained with either group normalization + weight standardization (GNWS) or batch normalization (BN). All approaches are trained only on 224×224 ImageNet examples.

Data source	BiT-ResNet-101x3 (ImageNet)	BiT-ResNet-101x3 (ImageNet-21k)	BiT-ResNet-101x3 (JFT)	CTX
Omniglot	$72.35 \pm 4.70\%$	$78.49 \pm 4.00\%$	$76.45 \pm 4.04\%$	$84.55 \pm 0.94\%$
Aircraft	$78.34 \pm 3.57\%$	$75.49 \pm 4.32\%$	$93.30 \pm 1.44\%$	$85.31 \pm 0.83\%$
Birds	$91.02 \pm 1.49\%$	$98.10 \pm 0.45\%$	$97.06 \pm 0.53\%$	$72.92 \pm 1.07\%$
DTD	$87.06 \pm 2.61\%$	$89.79 \pm 2.40\%$	$88.96 \pm 2.14\%$	$77.29 \pm 0.71\%$
QuickDraw	$65.08 \pm 4.13\%$	$69.16 \pm 3.79\%$	$71.27 \pm 3.77\%$	$73.29 \pm 0.78\%$
Fungi	$60.68 \pm 4.43\%$	$70.70 \pm 3.91\%$	$62.59 \pm 4.29\%$	$47.95 \pm 1.19\%$
Traffic Sign	$76.23 \pm 4.68\%$	$72.51 \pm 4.73\%$	$69.13 \pm 5.34\%$	$80.12 \pm 0.97\%$
MSCOCO	$69.74 \pm 2.69\%$	$76.07 \pm 2.26\%$	$76.36 \pm 2.23\%$	$51.39 \pm 1.06\%$
Caltech101	88.59 %	89.54 %	91.78 %	84.24 %
CIFAR100	58.35 %	78.08 %	76.32 %	37.51 %
Flowers102	81.88 %	99.09 %	99.33 %	81.75 %
Pets	89.97 %	92.00 %	95.45 %	70.88 %
Sun397	35.47 %	50.35 %	57.24 %	24.79 %
SVHN	79.23 %	69.08 %	66.47 %	67.22 %
EuroSAT	94.64 %	95.63 %	95.33 %	86.43 %
Resics45	76.71 %	80.77 %	85.76 %	67.65 %
Patch Camelyon	82.97 %	81.26 %	81.81 %	79.77 %
Retinopathy	73.85 %	75.27 %	72.02 %	35.48 %
CLEVR-count	70.73 %	66.75 %	61.54 %	27.89 %
CLEVR-dist	54.19 %	53.85 %	55.96 %	29.61 %
dSprites-loc	95.38 %	90.00 %	96.80 %	23.19 %
dSprites-ori	61.13 %	62.47 %	63.84 %	46.92 %
SmallNORB-azi	17.50 %	15.40 %	13.78 %	37.02 %
SmallNORB-elev	36.40 %	37.05 %	29.68 %	21.62 %
DMLab	45.58 %	45.37 %	48.22 %	31.92 %
KITTI-dist	82.24 %	78.45 %	78.62 %	54.34 %
MD-v2	75.06 %	78.79 %	79.39 %	71.60 %
VTAB (all)	68.04 %	70.02 %	70.55 %	50.46 %
VTAB (natural)	72.25 %	79.69 %	81.10 %	61.07 %
VTAB (specialized)	82.04 %	83.23 %	83.73 %	67.33 %
VTAB (structured)	57.89 %	56.17 %	56.05 %	34.06 %

Table 8: VTAB+MD accuracies for BiT-L learners trained on varying amounts of upstream data. CrossTransformers (CTX) accuracies are provided for context. All approaches are trained on 224×224 inputs.

Data source	BiT-ResNet-50 (JFT)	BiT-ResNet-50 (JFT, deduplicated)	BiT-ResNet-50 (JFT, class-ablated)
Omniglot	$69.37 \pm 4.42\%$	$69.89 \pm 4.71\%$	$69.10 \pm 4.72\%$
Aircraft	$87.13 \pm 2.28\%$	$86.27 \pm 2.25\%$	$73.09 \pm 3.76\%$
Birds	$92.50 \pm 1.24\%$	$92.59 \pm 1.16\%$	$79.22 \pm 2.92\%$
DTD	$87.43 \pm 2.05\%$	$87.48 \pm 2.21\%$	$87.72 \pm 2.14\%$
QuickDraw	$63.99 \pm 4.23\%$	$63.65 \pm 4.23\%$	$64.45 \pm 4.05\%$
Fungi	$56.03 \pm 4.22\%$	$56.48 \pm 4.47\%$	$54.94 \pm 4.53\%$
Traffic Sign	$66.21 \pm 4.94\%$	$66.13 \pm 5.03\%$	$63.79 \pm 4.98\%$
MSCOCO	$70.39 \pm 2.44\%$	$71.06 \pm 2.40\%$	$70.15 \pm 2.55\%$
MD-v2	74.13 %	74.19 %	70.31 %

Table 9: VTAB+MD accuracies for BiT-L learners trained on ablated JFT variants. The deduplicated variant of JFT removes all images that are found in MD-v2 test sources, and the class-ablated variant removes all images belonging to airplane-, birds-, and fungi-related classes. All approaches are trained on 224×224 inputs.