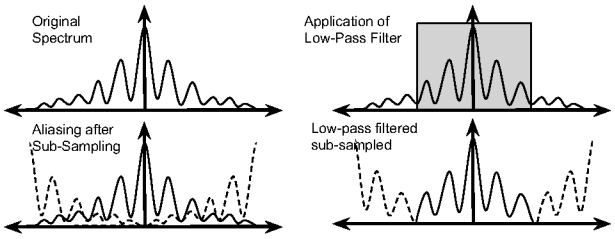


Figure 7: Subsampling without low-pass filtering causes the spectra to overlap and become corrupted. Left: after sub-sampling spectra could overlap, which is called aliasing; Right: subsampling preceded by low-pass filtering with an ideal low-pass filter prevents corruption.

Acknowledgement

Hugo Larochelle and Nicolas Le Roux are supported by Canada CIFAR AI Chairs.

A. The power spectral density of Binomial filters

This section reiterates the Fourier transform and its properties, and we provide the Fourier transforms for the filters used in our anti aliasing method. For a full exposition to the Fourier transform and frequency analysis please see [18]. In general the Fourier transform has both magnitude and phase components, our discussion focuses on the magnitude which is referred to as the power spectral density.

Figure 7 depicts the basic theory of frequency aliasing. When a signal is subsampled it's spectrum is replicated at distances inversely proportional to the sampling rate. Frequencies from these copies additively spill into the original signal, corrupting the original frequency components ("Aliasing after Sub-Sampling"). This aliasing can be prevented by the application of a low-pass filter, by which the lower-frequency components of the original signal can be preserved.

We use binomial filters as a low-pass filter because of their finite support size. Here we discuss the power spectral density of the 1D filter, which could be extended to the 2D case by an outer product. Examples of discrete binomial filters include $[1, 2, 1]$, $[1, 4, 6, 4, 1]$, etc, which can be generated from Pascal's triangle. Using the bracket $[\cdot]$ to index the signal, x, and defining the discrete Dirac delta function as δ, the filters of interest in our work are defined as:

$$
\begin{aligned}
x_1[n] &= \delta[n-1] + 2\delta[n] + \delta[n+1] \\
x_2[n] &= \delta[n-2] + 4\delta[n-1] + 6\delta[n] + 4\delta[n+1] + \\
&\quad \delta[n+2] \\
x_3[n] &= \delta[n-3] + 6\delta[n-2] + 15\delta[n-1] + 20\delta[n] + \\
&\quad 15\delta[n+1] + 6\delta[n+2] + \delta[n+3]
\end{aligned}
$$

Using the Fourier identities from Equations 1, we obtain the the corresponding signals in Fourier domain, denoting the angular frequency with w:

$$
\begin{aligned}
x_1[w] &= 2 + 2\cos(w) \\
x_2[w] &= 6 + 8\cos(w) + 2\cos(w) \\
x_3[w] &= 20 + 30\cos(w) + 12\cos(2w) + \cos(3w)
\end{aligned}
$$

Figure 8 shows the power spectral densities (magnitude of Fourier transform). This diagram highlights the trade-offs. A filter with larger support size, $k = 7$, attenuates more power at the cut-off frequency. However, more of the frequencies just below the cut-off frequency are also attenuated. A filter with smaller support size, $k = 3$, attenuates less power at the cut-off frequency, but maintains more information from the high frequencies just below the cut-off frequency.

$$
\begin{aligned}
\sum_{w=-\infty}^{\infty} x[n]e^{-jwn} &= x[w] && \text{(1a)} \\
\delta[n] &= 1 && \text{(1b)} \\
\delta[n-n_0] &= e^{-jwn_0} && \text{(1c)} \\
\delta[n-k] + \delta[n+k] &= 2\cos(wk) && \text{(1d)}
\end{aligned}
$$

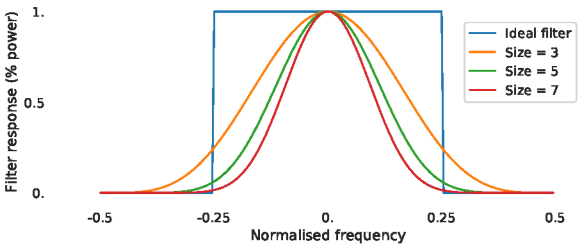
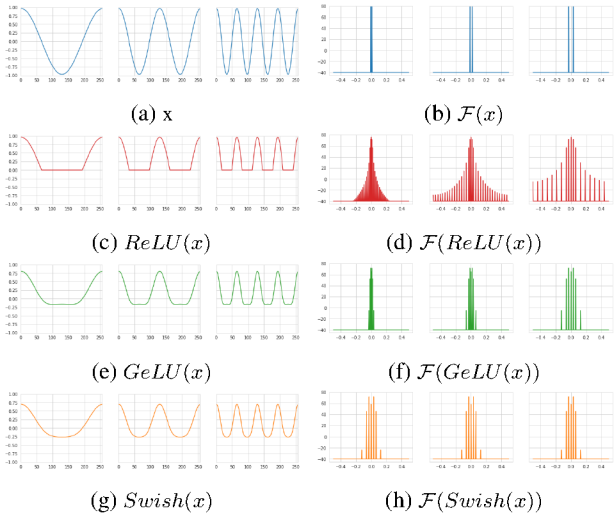


Figure 8: Power spectral density for filters used for anti-aliasing. The blue curve indicates an ideal filter for a sub-sampling with stride 2. However, as an ideal filter is too computationally expensive, we plot three alternatives of varying support size.

B. On Smooth activations

The optimal placement criteria presented in subsection 4.1 argues in favor of preserving the information encoded in high frequencies through layers that do not cause aliasing. Next, in subsection 4.2, our definition of the aliasing critical path claimed that the activation function non-linearity may produce high-frequency content. These arguments sustain the placement of low-pass filters in our models and its relation to activation functions (point-wise non-linearities) present in the original architecture.



(a) x (b) $\mathcal{F}(x)$ (c) $ReLU(x)$ (d) $\mathcal{F}(ReLU(x))$ (e) $GeLU(x)$ (f) $\mathcal{F}(GeLU(x))$ (g) $Swish(x)$ (h) $\mathcal{F}(Swish(x))$

Figure 9: Illustration of the spectrum changes caused by non-linear activation functions. First row shows three different inputs, each containing a different sinusoidal wave. Comparison between ReLU and smooth activations (GeLU and Swish) illustrates the impact on the spectra of the resulting signals. *Left:* spatial-domain representation. *Right:* frequency-domain representation.

This section illustrates the changes in spectral distribution caused by non-linear activation functions. Note that they introduce high-frequency components only if the inputs to the activation function span the non-linearity, i.e. for the activation functions considered here, inputs must be positive and negative.

Without loss of generality, Figure 9 illustrates the change in the power spectral density of 1D signals under different activation functions. The first row shows both spatial (x) and frequency domain ($\mathcal{F}$) representations of 1D sinusoidal waves, each containing a single frequency, i.e. this is our input signal without the application of any non-linearity. The following rows illustrate the spatial and frequency domain representations corresponding to the output of different non-linearities. The "elbow" spatial-domain characteristic (around zero) imposed by $ReLU$ is reflected in the resulting frequency-domain representation with the introduction of higher frequency components that were not present in the original input. The power of these new components decreases with their frequency. The results obtained by applying smooth-activation ($GeLU$ and $Swish$) to the same input signal also introduces new high frequency components, but with much faster decay. The proofs of the duality between spatial smoothness and frequency component decay can be found in [15] (page 41 shows that smoothness in time/spatial-domain implies decay in frequency-domain).

The use of smooth activations were previously shown to improve robustness to adversarial attacks [28]. The results of the ablation studies presented in sections 5.3 and 5.4 extend these findings to show their impact on a broader concept of o.o.d. robustness.

Figure 10 complements the results presented in Figure 5 by including the performance gains across different spectral bands for a model trained using smooth activation ("Swish") and also a model combining anti-aliasing with smooth activation and data augmentation ("Anti-aliasing + Rand Augmentation +Swish"). Similar to data-augmentation alone, smooth activations lead to improved performance mainly when the in lower bins are filtered, in contrast to the anti-aliasing which improves performance across the entire spectrum. The figure also shows that the combined model presents the best results at all frequency ranges, extending the benefits of using smooth activation functions, initially observed mainly in the lower frequencies, to the entire spectrum.

Figure 11 depicts a random sample of images and the corresponding images with various frequency bands removed. Note that when high frequency bands are filtered the images appear nearly indistinguishable from the originals.

C. EfficientNets

Large EfficientNet models are constructed by expanding a baseline model (EfficientNet-B0) in terms of model depth, model width, and input image resolution. For instance, EfficientNet-B0 adopts an input image of 224×224, EfficientNet-B1 scales it to 240×240 pixels and EfficientNet-B7 up to 600×600 pixels.

The results presented in this section were obtained with a EfficientNet-B0 in order to contrast the impact of aliasing on ResNet with EfficientNet models without introducing confounding effects related to input resolution. Table 4 shows that EfficientNet-B0 models also benefit from the introduction of small non-trainable low pass filters. This effect holds when training from ImageNet and using RandAugmentation. A comparison to ResNet-50 results shows that the impact of aliasing is smaller in EfficientNet-B0 – perhaps as a result of the fact that neural architecture search may have found an architecture which partially mitigates

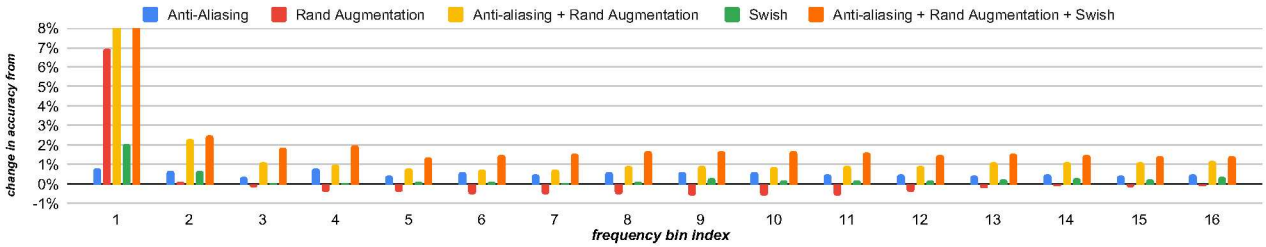


Figure 10: Illustration of the relative performance impact on pre-trained models when tested on images that have 1/16 of their spectral band removed. Our anti-aliased model performance is higher than the baseline in all spectral bands. The use of smooth activation functions alone have a larger impact in lower bands. The combined model (AA+DA+Swish) combines the advantages of all three. Note that this figure illustrates relative improvements to the baseline results taken under the same experiment. Baseline degradation curve is presented in Figure 5

Figure 11: Original and filtered test images pairs to illustrate the effect of the notch filter on each of the 16 frequency intervals (in order starting from lowest band). They are wrongly classified by the baseline model and correctly classified by the data-augmented model (bin 1) or the anti-aliased model (bin 2 - bin 16), but not by both.

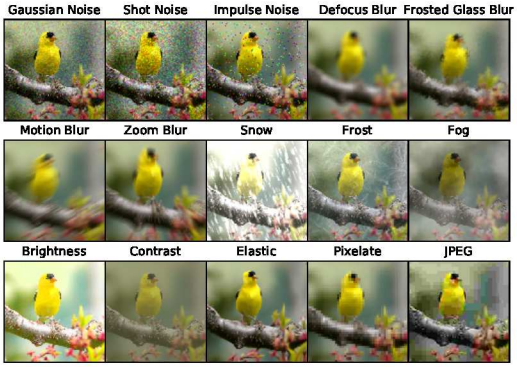


Figure 12: *Left*: ImageNet-C corruptions. *Right*: Samples from all 10 data sources included in the Meta-Dataset benchmark. Figure from Hendrycks and Dietterich [11].

these effects. EfficientNet models critical paths have a small 3×3 filter (Figure 4) that justifies the relative lower impact of anti-aliasing the EfficientNet model when compared to the impact observed on Resnet models. These results confirm our hypotheses that adding anti-aliasing to the critical paths is necessary in complement to the existing filters capacity (originally composed by 3×3 filters).

Model	Top-1 Acc.
Resnet-50	76.49 $\pm$ 0.06
Resnet-50 Anti-aliased	77.47 $\pm$ 0.12
Resnet-50 + Rand-Augmentation	77.38 $\pm$ 0.06
Resnet-50 Anti-Aliased + Rand-Augmentation	78.85 $\pm$ 0.02
EfficientNet-B0	76.40 $\pm$ 0.01
EfficientNet-B0 Anti-aliased (k=3)	76.65 $\pm$ 0.12
EfficientNet-B0 Anti-aliased (k=5)	76.58 $\pm$ 0.07
EfficientNet-B0 + Rand-Augmentation	76.98 $\pm$0.14
EfficientNet-B0 Anti-Aliased (k=3) + Rand-Augmentation	77.17 $\pm$0.10
EfficientNet-B0 Anti-Aliased (k=5) + Rand-Augmentation	77.12$\pm$ 0.08

Table 4: Comparison of the impact of aliasing on top-1 accuracy of EfficientNet versus Resnet models for input image with resolution 224×224. Table contains corresponding baselines, our anti-aliased versions and their combinations with data-augmentation. The values correspond to the mean accuracy over 3 runs with different seeds. Anti-aliased models present the best performance, but impact on Resnet-50 is larger than on EfficientNet-B0. Resnet-50 contains aliasing critical paths that lack a minimum size to represent low-pass filters, while EfficientNet-B0 don't have such critical bottlenecks.

Method	Blur filter location			Noise			Blur				Weather				Digital				mCE	Clean err.
	skip	max-pool	block-conv	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG		
ResNet-50 Published [11]				80	82	83	75	89	78	80	78	75	66	57	71	85	77	77	76.7	23.9
BlurPool [30]	✓	✓	✓	73	74	76	74	86	78	77	77	72	63	56	68	86	71	71	73.4	23.0
Ours	✓			68	70	70	72	87	78	75	73	69	50	53	66	81	84	68	70.9	22.9
Ours		✓		70	72	73	72	87	80	79	72	69	51	54	67	83	84	68	72.0	23.4
Ours			✓	70	72	72	72	88	80	77	73	69	50	53	66	81	84	67	71.6	23.2
Ours	✓		✓	69	71	72	72	87	79	76	72	68	50	52	64	81	84	68	70.9	**22.5**
Ours	✓	✓		70	71	73	72	80	79	75	71	67	51	53	67	82	85	68	71.2	22.9
Ours	✓	✓	✓	68	70	70	72	85	82	75	72	62	50	52	66	81	81	66	**70.0**	**22.5**

Table 5: Corruption Error (CE) on Imagenet-C corruptions, mCE, and Clean Error values when including our anti-aliasing variations. ResNet-50 and training for 90 epochs. Lower is better. We see that adding anti-aliasing decreases the errors on all corruptions except for Pixel and Blur. The errors were computed on the model achieving the median performance on ImageNet across 3 seeds. In our models, blur is not applied at the first convolutional layer (due to its large spatial support) and on the other sub-sampled modules it is applied at the precise location sustained by spectral analysis, ie. at the sub-sampling operation before its non linearities, as opposed to after as in [30].

Method	Blur filter placement			Noise			Blur				Weather				Digital				mCE	Clean err.
	skip	max-pool	block-conv	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG		
Baseline				70	72	72	71	87	79	76	73	69	51	53	66	81	81	67	71.2	22.6
RandAugment** [5]				60	58	60	70	90	76	80	70	67	44	50	57	80	86	64	67.4	22.8
Ours + RandAugm.**	✓			60	60	60	70	85	72	76	69	65	42	48	55	80	86	62	66.2	21.8
Ours + RandAugm.**	✓	✓		58	57	59	70	85	72	76	69	66	42	48	56	78	82	62	65.2	21.5
Ours + RandAugm.**	✓		✓	58	58	59	70	86	74	75	69	64	41	47	55	79	83	62	65.4	21.5
Ours + RandAugm.**	✓	✓	✓	59	58	61	70	84	75	76	69	65	41	48	55	80	82	61	65.5	21.6
Swish				71	72	74	69	88	80	76	74	69	51	54	68	81	80	67	71.6	22.4
Swish + Rand Augm.				61	61	62	69	88	73	78	69	67	42	49	55	81	87	63	66.9	21.9
Ours + Swish + Rand Augm.	✓			59	59	59	69	85	74	76	59	65	42	47	55	78	79	62	65.1	21.4
Ours + Swish + Rand Augm.	✓	✓		60	59	60	69	85	73	76	67	65	42	47	55	78	82	62	65.3	21.4
Ours + Swish + Rand Augm.	✓		✓	60	60	63	69	86	71	75	68	64	41	47	55	78	83	62	65.3	21.1
Ours + Swish + Rand Augm.	✓	✓	✓	60	59	61	69	85	71	75	68	64	41	47	55	78	81	61	64.9	21.2

Table 6: Corruption Error (CE), mCE, and Clean Error values when including our anti-aliasing variations on top of ResNet-50 and training for 180 epochs with data augmentation. Adding anti-aliasing leads to a lower error than all existing models with the exception of ANT. ANT uses adversarial training and has an extra generative network, is significantly more expensive to train, has a higher clean error and has comparable Corruption Error to our simple modification. The errors were computed on the model achieving the median performance on ImageNet across 3 seeds. In our models, anti-aliasing is applied before the non linearities, as opposed to after as in [30].

D. ImageNet-C: Robustness to Natural Corruptions

ImageNet-C [11] is a dataset used for evaluating the robustness of classifiers under natural image corruptions. It consists of the ImageNet validation set corrupted with 15 (plus four optional) types of natural corruptions under various severity levels (Figure 12 depicts ImageNet-C examples). These corruptions distort the distribution of the image spectra to varying degrees. In contrast to previously proposed methods [11, 12, 23], which achieve increased robustness (o.o.d.) at the cost of reducing i.i.d. performance (i.e. ImageNet validation without corruptions), we demonstrate that our method is the first to achieve state-of-the-art robustness without compromising accuracy on i.i.d performance. We show that our proposed architecture is complementary to data-augmentation and helps to achieve new state-of-the-art results on ImageNet-C. This result also suggests that anti-aliasing cannot be fully learned using existing augmentation strategies alone, without further architecture modifications.

Table 5 further detail the comparison between our anti-aliased model and Zhang's model [30] on ImageNet-C, when trained under the same number of epochs. It depicts the impact of aliasing in each of our model components (from Figure 3). Note that anti-aliasing the strided-skip connections alone already surpass Zhang's results in both ImageNet and ImageNet-C. Our combined model further improves the results using fewer and smaller filters.

Table Table 6 confirms the complementary impact of anti-alias in relation to data-augmentation and smooth activation functions in this o.o.d. setting. It contains baselines

Model	ImageNet	ImageNet-R	ImageNet-V2
Baseline	76.7	24.4	64.6
Zhang '19 [30]	77.2	24.1	65.0
Anti aliased (ours)	**77.5**	**25.2**	**65.2**

Table 7: Out of distribution generalization of anti aliased models compared to [30]

obtained with data-augmentation and/or the use of smooth activation functions alone, and contrasts them with the result obtained when also combining our anti-aliased model. The combination of the three improved both ImageNet and ImageNet-C results.

E. Additional experiments with Out-of-distribution Generalization

Here we report results on two additional o.o.d. generalization tasks for the models trained with anti aliasing filters. In Section 5.3 we analysed the robustness on ImageNet-C, which consists of synthetic perturbations. Here we make a 1-to-1 comparison with [30] on two datasets which represent natural robustness:

1. ImageNet-V2 [21] are images similar to those found in the ImageNet dataset. However this version was collected again in 2019. A high accuracy on this dataset indicates a better generalization to the new collection policy of ImageNet-V2, despite the original authors showing lower accuracy of most models on this dataset.

2. ImageNet-R [10] are renditions of the ImageNet classes, but in different styles such as sketches, paintings, or sculpture. Higher accuracy on this dataset indicates robustness to rendering method and image style.

In order to replicate [30]'s pipeline we also trained our models for 90 epochs. As discussed in Section 2, our models prevent aliasing in the feature maps of the neural networks, we expect our models to generalize better to these out-of-distribution datasets. Table 7 shows our results. The table shows that our anti aliased models have better o.o.d. generalization than [30], while maintaining similar accuracy on ImageNet. Specifically, anti aliasing improves from 24.2 to 25.2 on ImageNet-R, while [30] scores 24.1. We speculate this is due to many high frequency patterns in the renditions of ImageNet-R. Secondly, on ImageNet-V2 the baseline is 64.6, [30] scores 65.0, and our anti aliased model has the best accuracy at 65.2. While not the focus of our work, we show that the anti-aliasing approach shows higher accuracies on natural robustness.

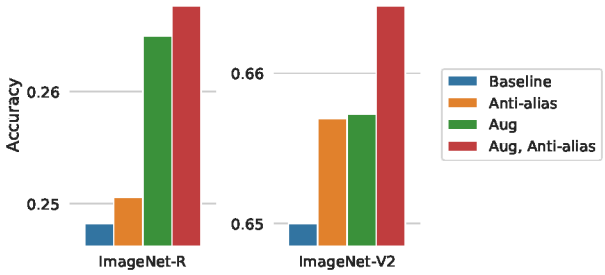


Figure 13: Our anti aliased models evaluated on two additional o.o.d. generalization tasks (described in Appendix E. Higher accuracy indicates better o.o.d. generalization. On all four datasets, we observe a complementary benefit of data augmentation with our anti aliasing method. *Aug* refers to a model trained with data augmentation.

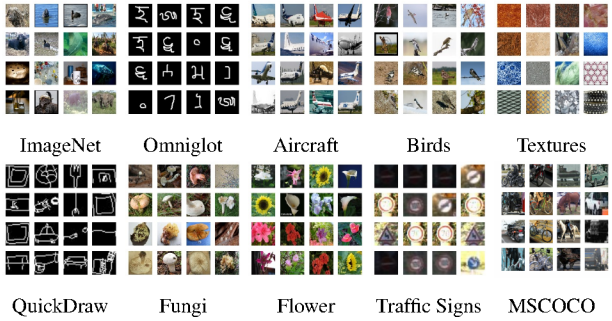


Figure 14: Samples from all 10 data sources included in the Meta-Dataset benchmark. Figures taken from Triantafillou et al. [26], respectively.

Secondly, we analyse the complementary effect of anti aliasing and data augmentation on o.o.d. generalization. In section 5.3 we observed a complementary effect of anti aliasing and training with data augmentation. Figure 13 shows the same models evaluated on ImageNet-R and ImageNet-V2. For both datasets, our proposed anti aliasing method brings an improvement over the baseline. Moreover, combined with data augmentation, the increase in accuracy (and thus robustness) is higher than the improvements of either method alone.

F. Few-shot classification, Meta-Dataset, and SUR

The objective behind few-shot classification is to create models which can learn on new problems with only a handful of labeled training examples. The evaluation procedure it prescribes is to form test *episodes* by subsampling classes from a held-out set of classes and sampling examples from those classes that are partitioned into a *support* (i.e. train-

Data source	Preprocessing	SUR*	Stride 1
Imagenet	↓↓↓	43.31	**46.40**
Omniglot	↓	97.28	97.11
Aircraft	↓↓↓	90.11	**90.88**
Birds	↓↓↓	70.87	**75.41**
Textures	↓↓↓	66.98	**71.17**
Quick Draw	↑↑	81.66	82.39
Fungi	↓↓↓	65.84	**70.95**
VGG Flower	↓↓↓	86.08	**88.60**
Average accuracy		75.27	**77.86**

Table 8: Subsampling in the first convolutional layer (i.e. before the first residual block) has a major impact on episodic validation performance for SUR trained on Meta-Dataset. Since the backbones are trained on 84×84 images, different datasets will require different amounts of upsampling or downsampling (upwards and downwards arrows in the *Preprocessing* column). For datasets requiring large amounts of downsampling (i.e. all datasets except Omniglot and QuickDraw), removing subsampling in the first convolutional layer (*Stride 1* column) shows clear benefits when compared with SUR's ResNet-18 implementation (*SUR** column).

ing) and a *query* (i.e. test) set of examples. The model is tasked with training on the support set and is evaluated on its query set accuracy, finally the query set accuracies of many test episodes are averaged to obtain a measure of model performance on new learning problems. A detailed description of the setup can be found in [26].

Meta-Dataset [26] is a large-scale few-shot classification benchmark that was introduced as a more realistic and challenging alternative to popular benchmarks such as mini-ImageNet [27]. While mini-ImageNet is constructed out of ImageNet classes (using 64, 16, and 20 classes to sample training, validation, and test episodes, respectively), Meta-Dataset is constructed out of many heterogeneous datasets whose classes are themselves partitioned into training, validation, and test sets of classes. Meta-Dataset, therefore, is a more challenging dataset in terms of robustness to distribution shift, which is compounded by the fact that two of its data sources (MSCOCO and Traffic Signs) are strictly reserved for test episodes (Figure 14 depicts Meta-Dataset examples).

SUR [6] tackles Meta-Dataset's domain heterogeneity by training separate backbones for each of the 8 data sources that define a training split of classes. Each backbone is trained to minimize classification error by sampling batches from its corresponding training set of classes. Hyperparameter selection is performed by evaluating on episodes sampled from each backbone's corresponding validation set of classes using a nearest centroid classifier (NCC) on top of the backbone embedding. Finally, during testing, all backbone embeddings are individually gated and concatenated to form a single representation used by a NCC. An optimization loop searches for the optimal gating coefficients using the loss on the support set, and predictions for the query set are made using the gating coefficients found by the optimization loop.

For our experiments we retrained SUR's 8 ResNet-18 backbones on their corresponding Meta-Dataset datasets using the original open source codebase and hyper-parameters. We note that SUR's codebase is affected by a bug that causes the examples of each class to be visited in a deterministic order when sampling episodes. This bug was fixed in our experiments. [3] This impacts both training and evaluation (Traffic Sign evaluation is particularly sensitive to the issue), which is why our reported baseline accuracies differ from those reported in the original SUR paper (the margin being wider for Traffic Signs).

SUR's preprocessing pipeline resizes the images of Meta-Dataset from their native resolutions to 84×84, using a bilinear interpolation [6]. In order to isolate the impact of this processing on aliasing artifacts, we look at its effect on episodic validation performance. As a reminder, validation for each backbone is performed using episodes sampled from its corresponding validation set of classes, which means that the combination of different backbones is eliminated as a potential confounding factor. Table 8 shows that backbones for which the input data needs heavy downsampling (represented by multiple downward pointing arrows in the *Preprocessing* column) benefit the most from the ablation of subsampling in the network's first convolutional layer.

Table 9 shows the impact of combining anti-aliasing with smooth GELU activation functions on Meta-Dataset's test episode accuracies. Adding low-pass filters on the skip connections yields an average accuracy of 74.80% (Anti-aliased skip + GELU). Including low-pass filters at all downsampling operations does not significantly improve average performance (74.82%). We highlight that an average improvement of 3.75% (absolute) was obtained (including a 2.73% improvement on out-of-domain tasks). Note that this was achieved with only minor changes to the architecture while using the default hyper-parameters.

[3] https://github.com/google-research/meta-dataset/issues/54

	SUR*	Anti-aliased		GELU	Anti-aliased + GELU	
Data source		$k = 3$	$k = 5$		$k = 3$	$k = 5$
Imagenet	53.77±1.10	57.32±1.13	58.46±1.08	56.81±1.11	59.91±1.06	**60.59±1.04**
Omniglot	95.81±0.36	96.11±0.36	95.87±0.36	96.29±0.32	96.06±0.33	**96.45±0.31**
Aircraft	87.47±0.49	89.22±0.43	89.54±0.45	88.36±0.56	90.76±0.46	90.34±0.50
Birds	72.44±0.98	78.70±0.87	77.86±0.82	72.81±0.89	78.52±0.81	79.96±0.79
Textures	68.96±0.77	71.29±0.90	72.08±0.78	72.59±0.80	74.40±0.73	74.98±0.80
QuickDraw	81.58±0.60	82.30±0.54	82.29±0.56	83.05±0.54	**84.01±0.54**	83.18±0.53
Fungi	65.67±1.01	69.87±0.93	71.07±0.98	67.91±1.03	72.88±0.92	73.50±0.89
VGG Flower	87.61±0.63	89.04±0.55	87.69±0.60	87.55±0.58	88.81±0.52	89.31±0.48
Traffic Signs	51.75±1.06	51.05±1.09	**55.52±1.03**	53.51±1.06	51.55±1.01	53.51±1.04
MSCOCO	48.95±1.10	48.30±1.09	49.61±1.05	49.71±1.05	50.28±1.04	**50.72±1.02**
MNIST	94.04±0.49	92.90±0.55	92.11±0.56	**95.42±0.43**	93.84±0.51	91.02±0.45
CIFAR10	62.45±0.91	66.22±0.84	66.06±0.85	63.20±1.01	**69.60±0.77**	69.29±0.71
CIFAR100	53.12±1.10	56.80±1.04	57.03±1.05	56.81±1.21	**61.58±1.04**	59.61±1.02
Average	71.05	73.02	73.48	72.62	74.79	**74.80**
Average (in-domain)	76.66	79.23	79.37	78.17	80.67	81.04
Average (out-of-domain)	62.06	63.07	64.07	63.73	65.37	**64.83**

Table 9: Evaluation of SUR models on 600 test episodes from Meta-Dataset. Columns: *SUR* shows baseline performance using original backbones. *Anti-aliased*: shows the effect of adding blur to strided-skip connections with different blur kernel sizes (k), *GELU*: replaces ReLU activations with GELU, *Anti-aliased + GELU*: combines the anti-aliased model with GELU activations. Meta-Dataset Anti-aliased backbones used stride 1 in the first convolutional layer with input size 84×84. Conclusions: adding blur at skip connections improves performance with or without GELU activations. GELU actions improve performance on their own. The best result is achieved by combining blur on skip connections with GELU activations.