MULLER: Multilayer Laplacian Resizer for Vision (Supplementary Material)

Zhengzhong Tu, Peyman Milanfar, Hossein Talebi
Google Research
zhengzhongtu@google.com

1. Overview

This supplementary document is organized as follows:

- We present detailed experimental settings and hyperparameters for image classification, object detection and segmentation, and image quality experiments in Sec. 2.
- Additional experimental results of MULLER resizer with respect to comparisons with previous works and the generalization are provided in Sec. 3.
- The discussion of the anti-aliasing effect as well as a comprehensive visualization are given in Sec. 4 and Sec. 5.

2. Experimental Settings

2.1. ImageNet Classification

We provide the experimental settings for both pre-training and fine-tuning MaxViT models on ImageNet-1K, detailed in Tab. 1. All the MaxViT variants employed similar hyperparameters except for that the the stochastic depth rate was tuned for each setting. It should be noted that we first pre-trained the backbone on ImageNet-1k/-21k/JFT with 300/90/14 epochs at a resolution of 224×224. Subsequently, the backbone was jointly fine-tuned with MULLER plugged-in at a higher resolution for an additional 30 epochs.

2.2. Object Detection and Segmentation

We evaluated MaxViT on the COCO2017 [3] object bounding box detection and instance segmentation task. The dataset comprises 118K training and 5K validation samples. All MaxViT backbones were pretrained on the ImageNet-1k dataset at a resolution of 224×224 following the same training protocol detailed in Sec. 2.1. These pretrained checkpoints were then used as the warm-up weights for fine-tuning on the detection and segmentation tasks. Note that for both tasks, the input images were resized to 896×896 before being fed into the MULLER resizer. The backbone was actually receiving a 640×640 resolution images for generating the box proposals. The training was conducted with a batch size of 256, using the AdamW [4] optimizer with learning rate of 3e-3, and stochastic depth of $0.3, 0.5, 0.8$ for MaxViT-T/S/B backbones, respectively.

2.3. Image Quality Assessment

We trained and evaluated the MaxViT model on the AVA benchmark [5]. Similar to [2, 7]. We pre-train MaxViT for resolutions: 224×224. Then we initialized the model with ImageNet-1K 224×224 pre-trained weights and fine-tune it with MULLER resizer. The weight and bias momentums are set to 0.9, and a dropout rate of 0.75 is applied on the last layer of the baseline network. We use an initial learning rate of 1e-3, exponentially decayed with decay factor 0.9 every 10 epochs. We set the stochastic depth rate to 0.5.

Hyperparameter	ImageNet-1K		ImageNet-21K		JFT-300M	
	Pre-train (MaxViT-T/S/B/L)	Fine-tune(+MULR) (MaxViT-T/S/B/L)	Pre-train (MaxViT-B/L/XL)	Fine-tune(+MULR) (MaxViT-B/L/XL)	Pre-train (MaxViT-B/L/XL)	Fine-tune(+MULR) (MaxViT-B/L/XL)
Stochastic depth	0.2/0.3/0.4/0.6	0.3/0.5/0.7/0.95	0.3/0.4/0.6	0.4/0.5/0.9	0.0/0.0/0.0	0.1/0.2/0.1
Center crop	True	False	True	False	True	False
RandAugment	2, 15	2, 15	2, 5	2, 15	2, 5	2, 15
Mixup alpha	0.8	0.8	None	None	None	None
Loss type	Softmax	Softmax	Sigmoid	Softmax	Sigmoid	Softmax
Label smoothing	0.1	0.1	0.0001	0.1	0	0.1
Train epochs	300	30	90	30	14	30
Train batch size	4096	512	4096	512	4096	512
Optimizer type	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Peak learning rate	3e-3	5e-5	1e-3	5e-5	1e-3	5e-5
Min learning rate	1e-5	5e-5	1e-5	5e-5	1e-5	5e-5
Warm-up	10K steps	None	5 epochs	None	20K steps	None
LR decay schedule	Cosine	None	Linear	None	Linear	None
Weight decay rate	0.05	1e-8	0.01	1e-8	0.01	1e-8
Gradient clip	1.0	1.0	1.0	1.0	1.0	1.0
EMA decay rate	None	0.9999	None	0.9999	None	0.9999

Table 1. **Detailed hyperparameters used in ImageNet-1K experiments.** Multiple values separated by '/' are for each model size respectively.

3. Additional Experimental Results

3.1. Comparisons to Previous Resizer

We compare the proposed MULLER resizer against the previous learned resizer with residual convolution blocks [8]. As shown in Tab. 2, fine-tuning with MULLER performs as effective as, and sometimes better than the previous heavier residual resizer. Furthermore, we note that MULLER is two orders-of-magnitude cheaper in inference cost (FLOPs), which further saves up to 52% training cost on TPUs, depending on the model size. Thus, MULLER is a promising 'green' machine learning model that can be easily integrated into various applications without incurring additional costs. Another benefit of MULLER over [8] is that MULLER is restricted to generating images that are more comprehensible to humans, despite being trained only for machine vision. This may be attributed to the bandpass design of the multilayer Laplacian filters employed in MULLER.

Model	Size	Infer cost (FLOPs)	Train cost (TPUv3 hrs)	top-1 accuracy
MaxViT-T	224	5.6B	-	83.62
+Residual $[8]_{512\to224}$	224	6.8B	2.8	83.93
+MULLER$_{512\to224}$	224	5.6B	1.9	83.95
MaxViT-S	224	11.7B	-	84.45
+Residual $[8]_{512\to224}$	224	12.9B	4.2	84.95
+MULLER$_{512\to224}$	224	11.7B	2	85.95
MaxViT-B	224	23.4B	-	84.95
+Residual $[8]_{512\to224}$	224	25.4B	6	85.48
+MULLER$_{512\to224}$	224	23.4B	3.5	85.58
MaxViT-L	224	43.9B	-	85.17
+Residual $[8]_{512\to224}$	224	45.1B	7.7	85.73
+MULLER$_{512\to224}$	224	43.9B	5.0	85.68

Table 2. **Performance comparison against previous residual resizer [8].**

3.2. Transferability Experiments.

We also examine the generalization ability of the learned resizer across different MaxViT model variants. Specifically, we take the learned resizer parameters from one MaxViT variant, and directly test it on another variant. As can be seen in Tab. 3, the learned resizer generalizes very well across different MaxViT model scales. The average top-1 accuracy drop is less than 0.06 when using different learned weights, indicating great transferrability of the MULLER resizer.

Model	MaxViT-T	MaxViT-S	MaxViT-B	MaxViT-L
$\text{MULLER}_{\text{M-T}}$	83.95	84.91	85.61	85.68
$\text{MULLER}_{\text{M-S}}$	83.96	84.95	85.61	85.68
$\text{MULLER}_{\text{M-B}}$	83.97	84.89	85.58	85.69
$\text{MULLER}_{\text{M-L}}$	83.95	84.91	85.61	85.68

Table 3. **Cross-model validation of the MULLER resizer for ImageNet-1K on MaxViT variants.** These values represent the top-1 accuracy of a given backbone tested with various MULLER resizers.

3.3. The Effect of Base Resize Method.

We conduct another ablation study to inspect the effects of the base resize method used inside MULLER. It is worth highlighting that this is the resizer used in MULLER, and it may or may not be different than the 'default resizier' mentioned in the main paper. Since we run all the experiments on TPU devices, we have found that only `bilinear` and `nearest` resizers are compilable. As demonstrated in Tab. 4, using nearest method as the base resizer yields similar performance as compared to the default bilinear method. We thus hypothesize that the choice of the base resize method used in MULLER does not significantly affect the performance of the model.

Model	Resize method	TPU compilable?	Top-1 acc.
MaxViT-B	Bilinear	Yes	85.58
	Nearest	Yes	85.54
	Bicubic	No	-
	Lanczos	No	-

Table 4. **Effects of base resize method used in MULLER.**

4. On Anti-Aliasing

We now investigate the effects of anti-aliasing on the input images to MULLER resizer. Our experiments reveal that while removing anti-aliasing does not affect the overall performance gain obtained by MULLER, the learned parameters may differ. As shown in Tab. 5, the learned parameters for each backbone have a slight shift in the weights and biases. However, these place no effects on the fine-tuning performances.

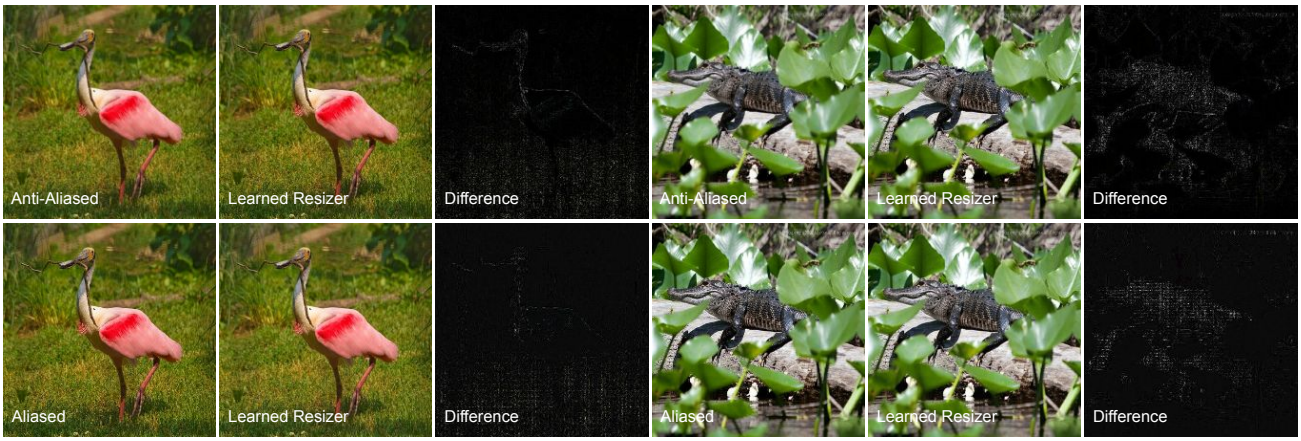
We do observe that anti-aliasing may impact the behavior of the learned resizer in terms of visualizations. For instance, as shown in Fig. 1, (a) when anti-aliasing is enabled for MaxViT-B, MULLER learns to enhance the contrast/details of the image to some extent; if the input image is aliased, nevertheless, MULLER learns to reduce the 'aliased effects'. In other words, the difference image displays some patterns similar to the aliasing effects in the resized image. (b) as for ResNet-50, it may be seen that MULLER learns to boost details even more for aliased inputs than the anti-aliased. Both effects have not been observed to significantly impact the performances, though.

5. Visualization

Figs. 2 and 3 illustrate some additional visualization results of the learned MULLER resizer for various backbones, including (a) EffNet-B0, (b) MobileNet-V2, (c) ResNet-50, and (d) MaxViT-B, arranged in ascending order of model complexity. A few of observations can be made: (1) On all the models, the MULLER resizer learns to boost the details/contrast of the image, albeit with varying degrees; (2) As evident from the performance gain of the vision models, the embedded information in the MULLER resized images is machine-friendly, and contributes to a more effective learning of the backbone; (3) Due

Anti-aliasing?	Model	α_1	β_1	α_2	β_2	top-1 acc.
Yes	EffNet-B0 [9]	1.715	0.088	-8.41	0.001	78.2
	MobileNet-v2 [6]	1.480	0.174	-5.25	-0.058	71.8
	ResNet-50 [1]	1.892	-0.014	-11.295	0.003	76.2
No	EffNet-B0 [9]	1.632	-0.014	-7.265	0.026	78.2
	MobileNet-v2 [6]	1.792	0.269	-7.514	-0.077	71.7
	ResNet-50 [1]	1.687	-0.039	-12.637	0.015	76.2

Table 5. The learned MULLER parameters for different backbone models train on ImageNet-1k. Top 3 rows show results using anti-aliased resizer while bottom 3 rows depict aliased resizing..

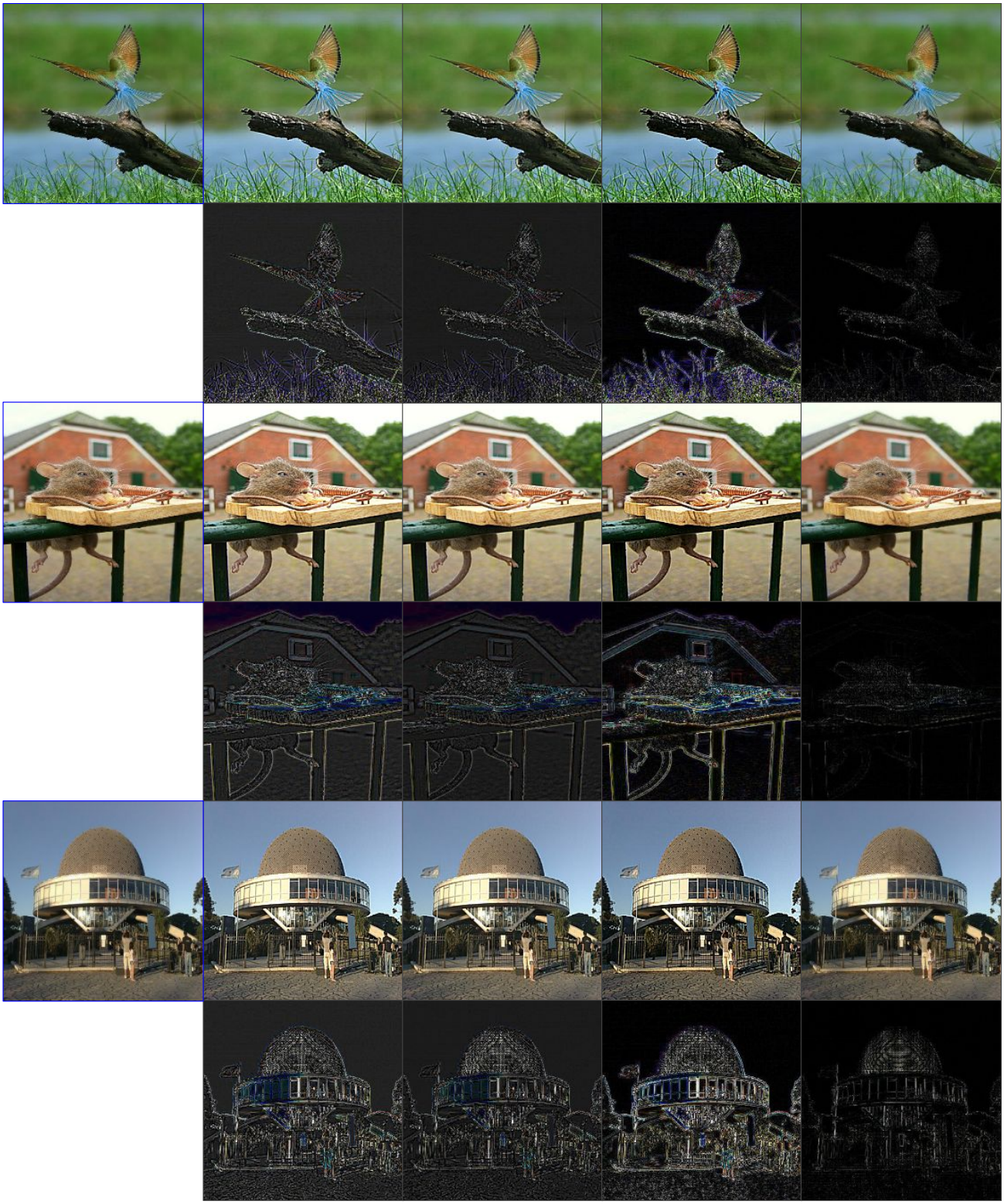


(a) Anti-aliased vs. aliased resize method for MaxViT-B

(b) Anti-aliased vs. aliased resize method for ResNet-50

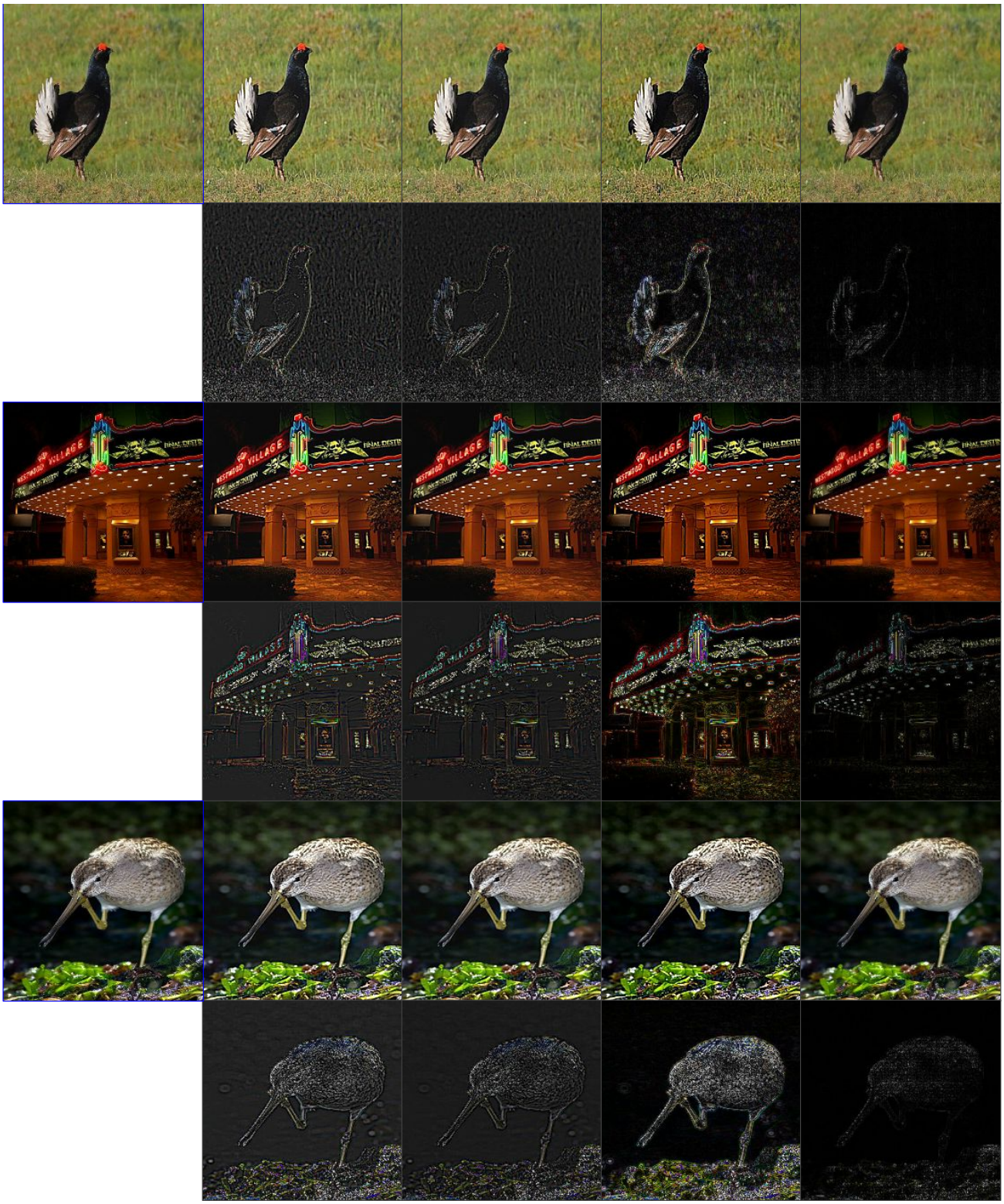
Figure 1. Visualization of the impact of anti-aliasing for the input image of MULLER. (a) shows examples for MaxViT-B, while (b) demonstrates those for ResNet-50.

to the highly regularized design of the resizer, the outputs of MULLER remain highly perceivable by human (in some cases even look perceptually superior), even though MULLER is purely trained for machine vision.



Default resized (a) EffNet-B0 (b) MobileNet-V2 (c) ResNet-50 (d) MaxViT-B

Figure 2. Visualizations of the MULLER resizer for (a) EffNet-B0, (b) MobileNet-V2, (c) ResNet-50, and (d) MaxViT-B. Here the default resizer is an anti-aliased resizer. Below each resized image shows the difference with the default resizer.



Default resized (a) EffNet-B0 (b) MobileNet-V2 (c) ResNet-50 (d) MaxViT-B

Figure 3. Visualizations of the MULLER resizer for (a) EffNet-B0, (b) MobileNet-V2, (c) ResNet-50, and (d) MaxViT-B. Here the default resizer is an anti-aliased resizer. Below each resized image shows the difference with the default resizer.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[2] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 1

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[5] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012. 1

[6] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 4

[7] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018. 1

[8] Hossein Talebi and Peyman Milanfar. Learning to resize images for computer vision tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 497–506, 2021. 2

[9] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 4