Supplementary Material for CLAD: A Contrastive Learning based Approach for Background Debiasing

Ke Wang
k.wang@epfl.ch
Harshitha Machiraju*[1,2]
harshitha.machiraju@epfl.ch
Oh-Hyeon Choung*[1]
ohhyeon.choung@gmail.com
Michael H. Herzog[1]
michael.herzog@epfl.ch
Pascal Frossard[2]
pascal.frossard@epfl.ch

[1] Laboratory of Psychophysics (LPSY) Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[2] Signal Processing Laboratory 4 (LTS4) Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

1 Effect of Negative Sample Dictionary

1.1 Ablation for Negative Sample Dictionary

To validate the effectiveness of negative sample dictionary, we conduct an ablation study. We originally create negative samples, which contain background associated with the anchor's foreground class. However, for this ablation, we create trivial negative samples, which contain background from random classes and are not necessarily matched with anchor's foreground class. These trivial negative samples are shared across all anchors.

Table 1 presents the performance comparison between CLAD and CLAD+ models, with their respective counterparts where trivial negative samples are used, denoted as CLAD (Trivial) and CLAD+ (Trivial).

Model	ORIGINAL↑	ONLY-FG↑	MIXED-RAND↑	MIXED-SAME↑	ONLY-BG-T↓	BG-GAP↓
CLAD	**95.9**	**93.8**	**87.5**	**90.1**	**31.3**	**2.6**
CLAD (Trivial)	95.5	93.0	85.3	88.9	37.2	3.6
CLAD+	**95.6**	94.6	**89.3**	**90.5**	**22.6**	**1.2**
CLAD+ (Trivial)	95.4	**94.7**	89.1	90.3	24.7	**1.2**

Table 1: Accuracy (%) comparison between CLAD, CLAD+ against their counterparts where trivial negative samples are used.

We observe that both CLAD and CLAD+ indeed perform better than their counterparts which use trivial negative samples.

1.2 Effect of Different Number of Negative Samples

In Fig. 1, we show the accuracy of using different queue sizes (which is also the number of negative samples) for the dictionary. We choose the size to be 32 as it is the best trade-off for both models.

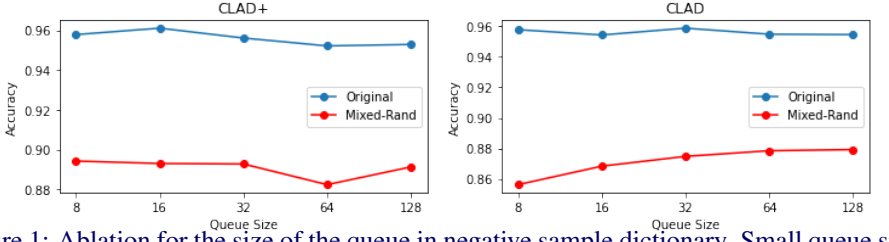

Figure 1: Ablation for the size of the queue in negative sample dictionary. Small queue sizes suffer from insufficient negative samples while large queue size suffers from deterioration of ORIGINAL accuracy and increased computational costs.

2 Foreground Segmentation and Scalability

In our experiments, following the Background Challenge dataset, we used 10 iterations of GrabCut [8] to segment the images' foreground using bounding box(bb) information. Instead of Grabcut which relies on bb, we tested using pre-trained U2-Net [7] to segment the foreground (Results in Table 2). We can see that the accuracy gap between the two methods on the ORIGINAL and MIXED-RAND dataset is within 1%, and results with U2-Net still beat all previous benchmarks on the Background Challenge. Since the performance does not change significantly, we can replace GrabCut with scalable methods like U2-Net, hence improving the scalability of our method.

In previous works in Table 2 of the main paper, [6, 9] also use foreground segmentation supervision, making them fair comparisons to ours.

Model	FG segmentation	Original ↑	MIXED-RAND ↑
Base (IN9)	-	**96.0**	73.4
CLAD+	GrabCut	95.6	**89.3**
	U2-Net	**96.0**	88.3
CLAD	GrabCut	95.9	87.5
	U2-Net	95.8	87.1

Table 2: Accuracy (%) comparison for using GrabCut and U2-Net as foreground segmentation methods.

3 Potential Foreground Positional Bias

The Background Challenge dataset may have centered foreground bias. Therefore, we use `FiveCrop` of PYTORCH to crop from the corners of the ImageNet-9 dataset and create foreground shift from the center. We report the averaged accuracy drop (Table 3). Our methods do not suffer from positional bias compared with baseline models. Contrastive learning penalizes positional shift bias because it enforces similarity between randomly augmented (main paper Sec.4.2) positive pairs.

	Base (IN9)	Base (MR)	CLAD+	CLAD
Accuracy drop (%) ↓	4.6	7.9	**4.1**	4.5

Table 3: Averaged accuracy drop (%) after corner cropping on Original dataset.

4 Mitigate Texture Bias

We show in this section how our method can be extended to texture biases. Previous works have shown that CNNs are biased towards local texture, instead of global shape [1, 2, 3, 4]. CNN's over-reliance on texture limits both its connection to human vision systems and its vulnerability to OOD data with texture-shape cue conflict [4]. Increasing CNN's shape bias would improve CNN's robustness towards a wide range of image distortions [3].

In this part, we show that the CLAD approach can be extended to other discriminative features. As an example, we show that it successfully reduces CNN's texture bias.

For the training scheme, we adopt the same approach as before and again experiment on the ImageNet-9 dataset, with the exception of how we generate contrastive pairs. We follow the basic idea that undesired discriminative feature (texture in this case) should be shared between negative sample pairs, while desired discriminative feature (shape) should be shared between positive sample pairs. In practice, when generating the cue-conflict images, we use the AdaIN [5] algorithm to modify the anchor's texture information. An example of the contrastive pairs used in our model (S-CLAD, S-CLAD+) for reducing texture bias is shown in Fig. 2.

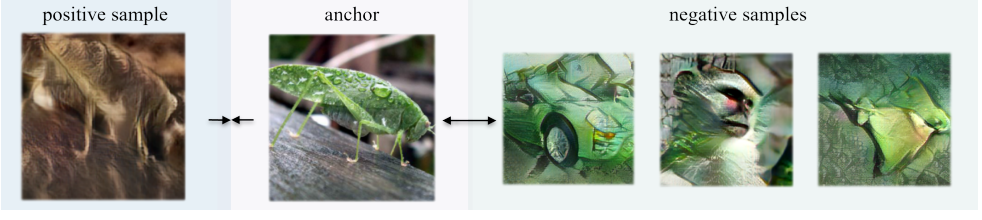

Figure 2: Example for sampling contrastive pairs for reducing texture bias

Datasets: We evaluate models' shape bias on two datasets: STYLIZED ImageNet-9 and ImageNet-9-SKETCH. We generate the STYLIZED ImageNet-9 using the same algorithm (AdaIN [5]) as the original Stylized ImageNet [3]. ImageNet-9-SKETCH is created by mapping the classes in ImageNet-Sketch [10] to classes in ImageNet-9. Models with high shape bias are expected to have better accuracy on these datasets, as the texture information in these datasets is either randomized or removed and hence provides no useful information on the class label.

The performance of S-CLAD+ and S-CLAD is compared against two baselines, `Base(IN9)` and `Base(SIN9)`, where the latter is a baseline model trained on stylized images from ImageNet-9 in a fully supervised setting. The results are presented in Table 4. We can see that, both S-CLAD and S-CLAD+ outperform the `Base(IN9)` baseline with a large margin on STYLIZED ImageNet-9 and ImageNet-9-SKETCH, indicating their texture bias is mitigated. Moreover, we again observe that there is almost no accuracy trade-off on the ORIGINAL ImageNet-9 for S-CLAD and S-CLAD+, whereas `Base(SIN9)` suffers from performance drop on the ORIGINAL ImageNet-9. Note that, no sketch images are included in S-CLAD+ and S-CLAD's training process, and they still have a performance gain of around 20% on the ImageNet-9-SKETCH dataset. This performance gain is because the model is

focusing more on shape information.

Model	ORIGINAL ImageNet-9	STYLIZED ImageNet-9	ImageNet-9-SKETCH
Base (IN9)	**96.0**	53.6	40.1
Base (SIN9)	91.5	75.1	58.0
S-CLAD	95.1	74.4	**61.0**
S-CLAD+	95.5	**76.7**	58.6

Table 4: Accuracy comparison for S-CLAD+ and S-CLAD, against `Base(IN9)` on three datasets. The accuracy on STYLIZED and SKETCH dataset indicates model's shape bias.

5 Cross-Evaluation

In this section, we cross-evalutate whether background-debiased models (CLAD and CLAD+) generalize better to the texture variation, and vise-versa. Firstly, we evaluate the shape bias of CLAD and CLAD+, compared against `Base(IN9)` on STYLIZED ImageNet-9 and ImageNet-9-SKETCH.

Model	ORIGINAL ImageNet-9	STYLIZED ImageNet-9	ImageNet-9-SKETCH
Base (IN9)	**96.0**	53.6	40.1
CLAD	95.9	**54.3**	39.7
CLAD+	95.6	53.7	**41.2**

Table 5: Evaluation of shape bias for CLAD, CLAD+, compared against `Base(IN9)`.

The results show a minor improvement by CLAD and CLAD+ on STYLIZED ImageNet-9 and ImageNet-9-SKETCH respectively. This can explained due to the fact that by removing background bias, we focus more on both the foreground shape and texture information. Hence, our model may still use the texture information from the foreground for classification along with its shape. Thus, its performance on datasets which transform the texture of the foreground and background together is not improved drastically. This also implies that background debiasing alone is not sufficient for texture debiasing.

Model	ORIGINAL↑	ONLY-FG↑	MIXED-RAND↑	MIXED-SAME↑	ONLY-BG-T↓	BG-GAP↓
Base (IN9)	**96.0**	86.0	73.4	87.5	42.9	14.1
S-CLAD	95.0	**87.5**	**78.9**	**89.3**	40.9	**10.4**
S-CLAD+	95.5	87.4	78.1	88.5	**38.1**	**10.4**

Table 6: Evaluation of S-CLAD, S-CLAD+ on Background Challenge, compared against `Base(IN9)`.

We then test S-CLAD and S-CLAD+ on the Background Challenge datasets [1]. As shown in Table. 6, we find that inducing shape bias helps mitigate background bias to some extent. This can be explained by the increased focus on shape of the object which is usually in the foreground, while also ignoring the background information. However, texture-debiased model, S-CLAD and S-CLAD+ alone are not sufficient to reproduce our state of the art results we had from CLAD and CLAD+ on Background Challenge.

References

[1] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 2018.

[2] Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. *International Conference on Learning Representations*, 2019.

[3] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*, 2019.

[4] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 2020.

[5] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *IEEE international conference on computer vision*, 2017.

[6] Sangjun Lee, Inwoo Hwang, Gi-Cheon Kang, and Byoung-Tak Zhang. Improving robustness to texture bias via shape-focused augmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[7] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 2020.

[8] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 2004.

[9] Chaitanya K Ryali, David J Schwab, and Ari S Morcos. Leveraging background augmentations to encourage semantic focus in self-supervised contrastive learning. *arXiv preprint arXiv:2103.12719*, 2021.

[10] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 2019.

[11] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *International Conference on Learning Representations*, 2020.