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

Abstract

Convolutional neural networks (CNNs) have achieved superhuman performance in multiple vision tasks, especially image classification. However, unlike humans, CNNs leverage spurious features, such as background information to make decisions. This tendency creates different problems in terms of robustness or weak generalization performance. Through our work, we introduce a contrastive learning-based approach (CLAD) to mitigate the background bias in CNNs. CLAD encourages semantic focus on object foregrounds and penalizes learning features from irrelevant backgrounds. Our method also introduces an efficient way of sampling negative samples. We achieve state-of-the-art results on the Background Challenge dataset, outperforming the previous benchmark with a margin of 4.1%. Our paper shows how CLAD serves as a proof of concept for debiasing of spurious features, such as background and texture (in supplementary material).
[1]

1 Introduction

CNNs have achieved superhuman performance on various computer vision tasks such as segmentation [35], classification [9, 10, 33], object detection[38], etc. However, it has been observed that CNNs have a different understanding of images in contrast to humans [5]. Specifically, in the case of classification, it has been observed that CNNs can be biased towards the background information instead of the foreground object [1, 2, 29, 36, 40], high-frequency components [20, 25, 37], and textures rather than shapes [6, 12]. In particular, Kai et al.[36] showed that CNNs tend to correlate class labels heavily with background information. Further, they showed that, when the foreground object is removed, CNNs still perform

[1]Code: https://github.com/wangke97/CLAD.

surprisingly well solely in the presence of the background of the image. The authors created the Background Challenge [36] which measures models' robustness to various changes the background. They further showed that most state-of-the-art image classification models exhibit a poor generalization ability in this challenge due to large background bias. These biases lead to an over-dependence of the model on irrelevant/spurious features. Further, such biases can be exploited to fool the classifiers by simply altering the background of the object [23] or adding different textures to the image [6].

To mitigate such biases, conventional data augmentation is often used, wherein the model is exposed to additional training data, in order to decorrelate the spurious features and the class label. However, to completely eliminate bias and prevent memorization (overfitting) of data, the model usually requires a very large amount of data for augmentation. Also, previous works [34, 39] have shown that conventional data augmentation is insufficient to discard spurious features and remains susceptible to small changes. Therefore, an effective data augmentation method has to be applied to ensure that the best features are extracted during training while introducing minimal computational overhead.
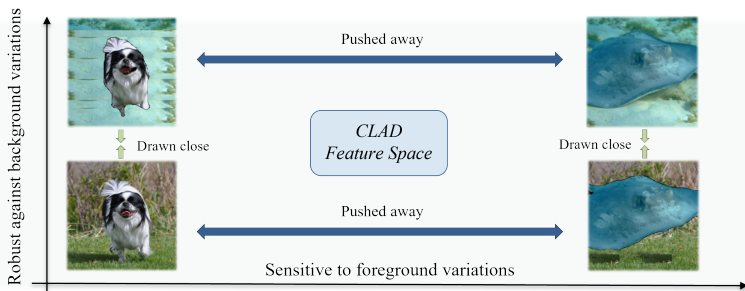

Figure 1: CLAD learns feature space which is robust against background variations and sensitive to foreground features.

Here, we propose a **C**ontrastive **L**earning based **A**pproach for background **D**ebiasing (CLAD) model, where contrastive learning (CL) is introduced to mitigate the biases more effectively. In contrastive learning, for each data point (anchor), both positive samples (sharing anchor's distribution) and negative samples (carries different information as anchor) are generated. Then, CL minimizes the distances between the anchor and positive samples and maximizes the distances between the anchor and negative samples in feature space. To this end, in the latent embedding, attributes that belong to the same distribution (relevant features, e.g., foregrounds) are aggregated together, while unwanted biases (e.g., backgrounds) are separated from the anchor. Hence, our method, CLAD, uses contrastive learning to learn a background-robust feature space, by carefully constructing the positive and negative samples. Positive samples are generated by changing the anchor's background. The negative samples, on the other hand, contain distinctive foreground information but similar background as the anchor (Fig. 2). Moreover, instead of generating negative samples, we introduce a novel mechanism to sample negative samples without introducing extra costs, where we sample negative pairs during the training process from generated positive samples while ensuring the sampled negative samples share similar background information as anchor. Thus, our novel method allows for both the scalability of negative samples as well as having similar background as anchor.

We show that, our CLAD model **outperforms** the state-of-the-art methods on the Back-

ground Challenge dataset [56], while it has almost no accuracy drop on original images. It samples negative samples effectively without introducing heavy computational costs. Especially, CLAD outperforms on the random-background dataset (MIXED-RAND) by a margin of 4.1%, while all the other state-of-the-art methods showed a major performance drop. We also show that CLAD can be applied to mitigate the influence of other discriminative features apart from background, like object texture (in supplementary material), while improving model's shape bias.

2 Related Work

Feature biases are thought to happen because of the data memorization (overfitting) and are exacerbated when training the over-parameterized models [14]. One effective way to mitigate these problems is to augment with samples emphasizing desirable features instead of irrelevant spurious ones. In background-biased settings, Kai et al.[56] showed that training models on images with random unrelated backgrounds for a given foreground helped reduce the background bias of the model. However, this also significantly reduced performance on the original dataset (Table 2). Further, as mentioned in the previous section, conventional data augmentation is not optimal for debiasing; hence, we rather look at contrastive learning.

Contrastive Learning (CL) [9] helps learn robust feature spaces that are close across a data distribution and attributes that set apart a data distribution from another. CL has shown great promise in self-supervised regimes [4, 7, 11] while recently, it has also been applied to the supervised learning domain and achieved promising results [15, 18, 19]. CL has been used in a self-supervised manner to help debias models [17, 22, 27, 52]. In the fully supervised learning domain, previous works have shown that utilizing contrastive loss as an auxiliary loss can encourage learning more robust features with higher generalization abilities through careful contrastive pair construction [18, 19]. To the best of our knowledge, we are the first to leverage contrastive learning as an auxiliary loss to improve the model's background robustness in a fully-supervised setting.

3 Methodology

In this section, we go through the contrastive learning framework and then introduce our background-debiased contrastive pair sampling strategy, and finally present our overall learning framework.

3.1 Contrastive Learning

We use the popular InfoNCE [8, 24] loss as our contrastive loss term. This loss function can be viewed as an (N+1)-way cross-entropy classification loss to distinguish between one positive sample and N negative samples, and is written as:

$$\mathcal{L}_{con} = -log\left[\frac{e^{s(x,x^+)/\tau}}{e^{s(x,x^+)/\tau} + \sum_{i=1}^{N} e^{s(x,x_i^-)/\tau}}\right] \quad (1)$$

where $s(x_1, x_2) = (x_1 \cdot x_2)/(\|x_1\| \|x_2\|)$ is the cosine similarity function and τ is the temperature parameter; x, x^+, x_i^- represent the feature representations for the anchor, the positive

sample and the multiple negative samples, respectively. It brings positive sample pairs closer in the feature space, while it pushes the anchor apart from negative samples.

3.2 Background-debiased Sampling

One crucial contribution of CLAD is an efficient sampling approach for contrastive pairs which are harder to discrminate from the anchor. Conventionally, in contrastive learning, positive samples are obtained by applying a combination of different data augmentations to the anchor. Negative samples, on the other hand, come from views of other images (see Fig. 2 (a)). However, such sampling of contrastive pairs would lead to poor robustness on backgrounds due to two reasons: 1) increasing feature similarity between positive pairs would simultaneously encourage background bias due to their shared background information; 2) likewise, as negative samples carry different background information compared to the anchor, minimizing feature similarity between negative pairs would increase the model's sensitivity to background variations.

These problems are solved in CLAD's background-debiased contrastive pair sampling approach, where background information is no longer shared between positive pairs, and negative pairs share similar background information, as shown in Fig. 2 (b). The contrastive pairs are created as follows:

Positive Samples are created by replacing the background of the anchor with a different-class background (chosen randomly). Following the method in Background Challenge dataset [36], we use GrabCut [26] to separate the foreground and background of a given anchor image (see supplementary material for details). The foreground of the anchor is then placed in a background found in another random class (other than the anchor class).

Negative Sample It is crucial to have a large number of negative samples in contrastive learning [4, 11]. However, using the same method to create positive samples, i.e., replacing the foreground of the anchor image instead and keeping the background, needs to be repeated many times to create multiple negative samples. This leads to a high computational cost which linearly scales the cost per batch by the number of negative samples. To solve this issue, we introduce a negative sample dictionary.
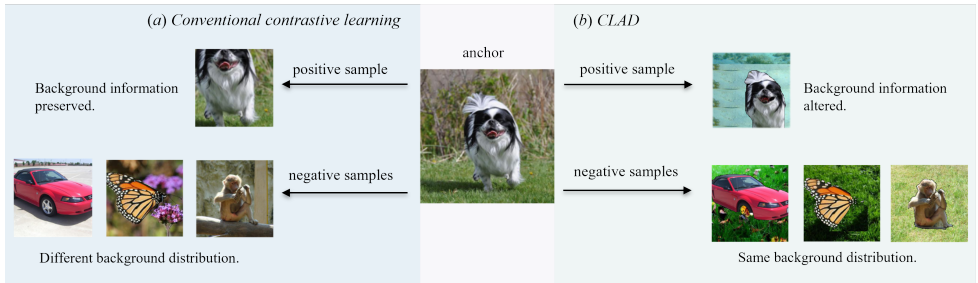

Figure 2: Contrastive pair sampling strategies, (a) used in conventional contrastive learning, (b) CLAD's background-debiased sampling strategy

We define our negative sample dictionary as a dictionary with queues for each class, containing the latent representation for each negative sample. Each queue, has samples whose background belongs to the class represented by the queue. The size of each queue is the same as the number of negative samples (N). In each batch, we use the generated

positive samples to update the queue. The old samples are dequeued (deleted) when new samples are enqueued (added) to the queue following a first in, first out order. Therefore, the negative samples are reused until they get replaced in the queue.

This differs from the commonly used memory bank [☐] for storing negative samples in two ways:

- It only stores features for background-augmented images where the foreground and background classes are decoupled.
- As a dictionary, it contains keys for **background labels** of the stored samples. Samples are stored in the queue whose key corresponds to their **background labels**.

We illustrate the mechanism in Fig. 3. The dictionary contains the keys of the background label, and we show two examples in the Figure. In the example for updating the dictionary with generated samples, the sample has a background of `Fish` class, so it will enter the queue within the `Fish` key (the foreground label is ignored in this process). The other example shows the sampling process for negative samples from the dictionary: the anchor is an image from `Dog` class; hence we draw all samples in the queue within the `Dog` key in the dictionary.
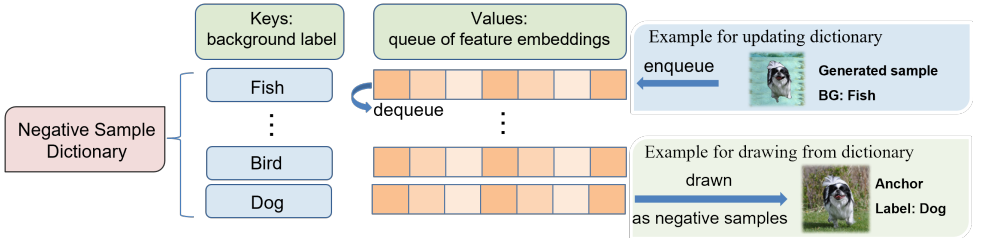

Figure 3: Illustration of using dictionary to store negative sample candidates.

Using the negative sample dictionary guarantees that similar background information is shared between negative pairs simultaneously. Hence, our method provides a memory-efficient way of scaling negative samples.

3.3 Training Objective

The overall loss function is composed of two terms: the conventional supervised classification loss (for learning distinguishable features) and contrastive loss (for improving background robustness). After we generate positive and negative sample pairs (as described in Sec 3.2), we calculate the contrastive loss using the InfoNCE loss function. To enforce the correct classification of the positive samples, we can optionally include a classification loss for such samples and refer to the model with this additional loss term as CLAD+. For the supervised classification loss, we use the conventional cross-entropy loss. Specifically, the overall loss for CLAD can be written as:

$$\mathcal{L}_{CLAD} = \mathcal{L}_{class}(x) + \lambda * \mathcal{L}_{con}(x, x^+, x^-) \tag{2}$$

For CLAD+, the loss is written as:

$$\mathcal{L}_{CLAD+} = \mathcal{L}_{class}(x) + \mathcal{L}_{class}(x^+) + \lambda * \mathcal{L}_{con}(x, x^+, x^-) \tag{3}$$

Here, λ is a hyperparameter for the weight that controls the importance of the contrastive term $\mathcal{L}_{con}$. Its magnitude controls the degree of background robustness learned by the model.

3.4 Training

As illustrated in Fig. 4, for each batch, we generate positive samples. Then, the generated positive samples are used to update the negative sample dictionary. The classification loss is calculated for the anchor (and also for the positive sample for CLAD+). When calculating the contrastive loss, the negative samples are drawn accordingly from the negative sample dictionary based on the label of each anchor. The contrastive loss is finally calculated based on feature representations for the anchors, positive and negative samples.
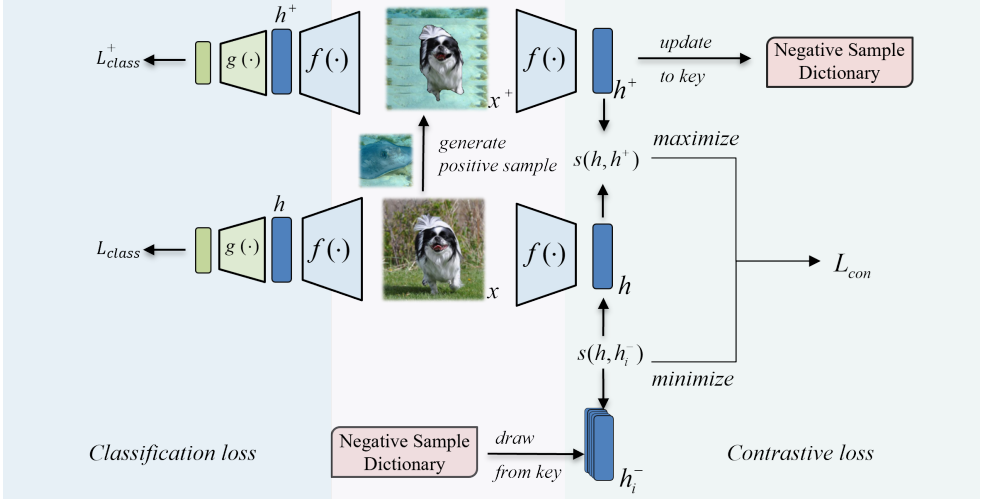

Figure 4: Illustration of the proposed supervised learning with contrastive learning approach.

4 Experiment

In this section, we present the results for CLAD and CLAD+ on the Background Challenge dataset [36].

4.1 Challenge Description

Kai et al.[36] initiated the Background Challenge [36] dataset in 2020. The dataset aggregates a subset of images in ImageNet based on WordNet hierarchy [21] into 9 classes, creating the ImageNet-9 dataset. Several variations are made on the images' background or foreground in original ImageNet-9, as summarized in Table 1.

Dataset	Foreground	Background	Summary
ORIGINAL	Original	Original	Original unaltered images
ONLY-FG	Original	None (Black)	Images with only the foreground (background removed)
ONLY-BG-T	None	Original	Images with only the background
MIXED-RAND	Original	Completely Random	Images with a random background
MIXED-SAME	Original	Random-same-class	Images with a background from the same class

Table 1: Description of the variations of ImageNet-9[36]

The goal of the Background Challenge is to achieve high accuracy on the MIXED-RAND dataset, where the background class is selected randomly and provides no information on image label. Intuitively, models with high background bias would suffer from low accuracy on this dataset. Additionally, the challenge also defines a metric to quantify the background bias: BG-GAP, which is defined as the accuracy gap between the MIXED-SAME and MIXED-RAND datasets. The BG-GAP represents the performance drop due to background class signal change [36], or more intuitively, how much accuracy is actually gained by background bias.

4.2 Experimental Settings

We adopt a ImageNet-pretrained ResNet-50 as our backbone [36]. Adam [16] is used as the optimizer with default settings ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) and no weight decay is used. The total number of training epochs is 60, and the batch size is 64. The learning rate is set to be $1e^{-3}$ and decays to $1e^{-4}$ after 20 epochs. After trial and error, the hyperparameter λ for the weight of the contrastive loss is set to 1 (ablation in 4.4) and the temperature parameter τ is set to 0.2. Data augmentations, including Random Resized Crop, Random Horizontal Flip, and Color Jitter, are used in our experiments. Note that for the generated positive samples, these conventional data augmentations are applied after the background augmentation. For each anchor, we construct one positive sample and draw 32 negative samples (details in supplementary material) from the negative sample dictionary. In this section, we evaluate the performance of CLAD on the Background Challenge dataset. For comparison, we compare the performance of CLAD to three baselines, which are trained in conventional, fully supervised settings, which include:

- `Base(IN)` ImageNet-trained ResNet-50 with prediction mapped to ImageNet-9.
- `Base(IN9)` ResNet-50 trained on ORIGINAL with a fully supervised setting.
- `Base(MR)` ResNet-50 trained on MIXED-RAND with a fully supervised setting.

In addition we also compare to previous works on Background challenge dataset, the results of which are presented in Table 2.

4.3 Accuracy

CLAD and CLAD+ do not suffer any accuracy trade-off on the ORIGINAL dataset compared to the baseline models (0.4% and 0.1% drop correspondingly). Our method outperforms all previous benchmarks by a large margin (4.1% for CLAD+ and 2.3% for CLAD) on MIXED-RAND dataset, which is the most important indicator for the model's generalization ability to varying-background images.

It is possible to have a very small BG-GAP as well as very low accuracy on both the MIXED-SAME and MIXED-RAND datasets. However, that would not be reflective of the background bias or generalization ability of the model. Hence, we need high performance on both datasets along with a smaller gap between them, to have less background bias. We plot the accuracies of these datasets in Fig. 5, wherein models that lie closer to the identity line have lower background bias. Additionally, the further right from the model's line, the higher its bias. We can see from the Figure that our models CLAD and CLAD+ have the best performance among all the models.
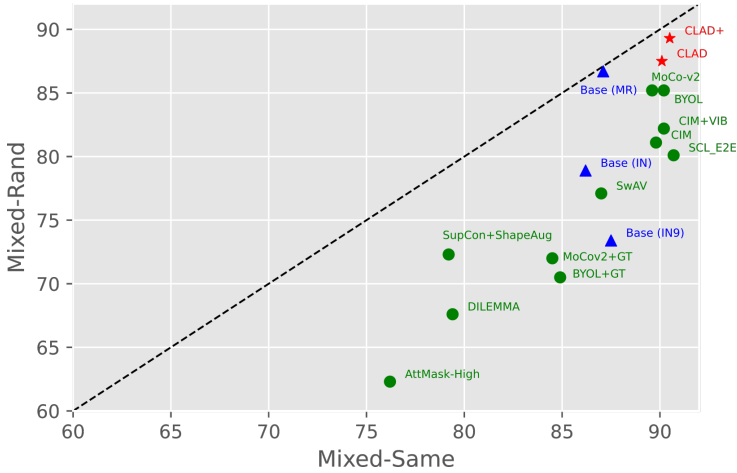
Figure 5: Model performances on MIXED-SAME (x-axis) and MIXED-RAND (y-axis) data. Models closer to the Identity dashed line has lower background bias.

Type	Model	ORIGINAL ↑	ONLY-FG ↑	MIXED-RAND ↑	MIXED-SAME ↑	ONLY-BG-T ↓	BG-GAP ↓
	Base (IN) [⬛]	96.2	-	76.3	82.3	17.8	6.0
Baselines	Base (IN9)	96.0	86.0	73.4	87.5	42.9	14.1
	Base (MR)	88.4	89.5	86.7	87.1	12.8	0.4
	CIM [⬛]	97.7	-	81.1	89.8	-	8.8
	SCL_E2E [⬛]	**98.2**	-	80.1	**90.7**	-	10.6
	CIM+VIB [⬛]	97.9	-	82.2	90.2	-	8.0
	SupCon+ShapeAug[⬛]	-	-	72.3	79.2	-	6.89
	MoCo-v2 (BG Swaps)[⬛]	95.2	87.5	85.2	89.6	**11.4**	4.4
Others	BYOL (BG Random)[⬛]	96.1	88.3	85.2	90.2	12.9	5.0
	SwAV (BG RM)[⬛]	95.3	86.8	77.1	87.0	18.2	9.9
	AttMask-High[⬛]	89.8	75.2	62.3	76.2	15.3	9.9
	MoCov2+GT[⬛]	89.7	72,7	72.0	84.5	40.1	12.5
	BYOL+GT[⬛]	91.0	72.6	70.5	84.9	41.2	14.4
	DILEMMA[⬛]	91.8	77.8	67.6	79.4	**9.3**	10.2
Ours	CLAD+	95.6	**94.6**	**89.3**	90.5	22.6	**1.2**
	CLAD	95.9	**93.8**	**87.5**	90.1	31.3	**2.6**

Table 2: Accuracy (%) comparison between CLAD, CLAD+ against baselines and benchmarks on the Background Challenge. '-' represents value missing in the references. Note that the models in this table are trained with different sizes of dataset and level of supervision, in this case the BG-GAP is the fairest comparison across all models indicating background bias.

4.4 Analysis

Feature Consistency We estimate the percentage of encoded foreground information by calculating the features' cosine similarity between image pairs sharing the same foreground. This metric can also be intuitively reflect as how much of the features are extracted from the foreground. We also define a more direct metric, decision consistency, which summarizes the fraction of consistent decisions after background change. This can be expressed as $\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\arg\max g(x_i) = \arg\max g(\hat{x}_i))$, where, $g(.)$ is the classifier, $(x_i, \hat{x}_i)$ represent image pairs with same foreground but different background. The higher the decision consistency, the smaller the effect of background changes on the models' decisions. For details, see Table

3.

Model	Feature Similarity	Decision Consistency
Base (IN9)	0.795	0.800
Base (MR)	0.864	0.864
CLAD+	**0.920**	**0.969**
CLAD	0.914	0.915

Table 3: Feature similarity and decision consistency between ORIGINAL and MIXED-RAND datasets. CLAD and CLAD+ can extract features over 90% similar to the extracted features before background variation, meaning that a large amount of the features they learned from the original images are from the foreground, well explaining their performance on the background challenge dataset.

Interpretability: Saliency map provides intuitive illustration for models' areas of focus in images. Fig. 6 illustrates the SmoothGrad[30] saliency maps of the CLAD and CLAD+, compared with two baseline models. It shows that the saliency maps for CLAD+ and CLAD focus more on the foreground object with a much cleaner saliency map than `Base(IN9)` and even `Base(MR)`. An interesting observation is that, the saliency map of `Base(IN9)` and `Base(MR)` on the `wolf` image (second row) shows these baseline models rely on background snow for identifying wolf, which is a well-known example for CNN's background bias [51]. CLAD+ and CLAD are able to identify `wolf` while ignoring background snow, relatively better than base models.
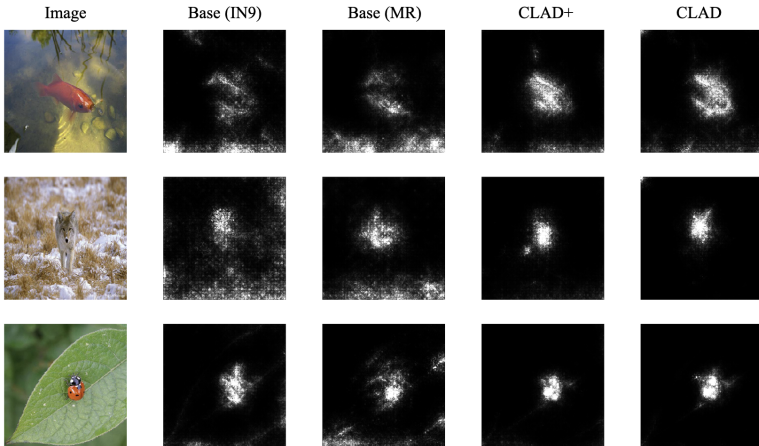

Figure 6: Saliency maps of CLAD+ and CLAD, compared with two baseline models.

Importance of Contrastive Loss: The hyper-parameter λ is the weight of the contrastive loss term in our overall loss function. In Fig. 7 we show that the magnitude of λ determines the background robustness of the CLAD model. This Figure presents the varying accuracy on MIXED-RAND dataset, as well as ORIGINAL IMAGENET-9, with increasing λ. The CLAD models do not have any performance deterioration when the contrastive loss is introduced with equal weight as the classification loss. Its performance on ORIGINAL dataset remains the same while increasing on MIXED-RAND. However, if the value of λ is further increased, i.e., the contrastive loss becomes more important than the supervised losses ($\lambda > 1$), then there is a performance deterioration. This indicates that we need a balanced mix of both supervised and contrastive losses for ideal performance.
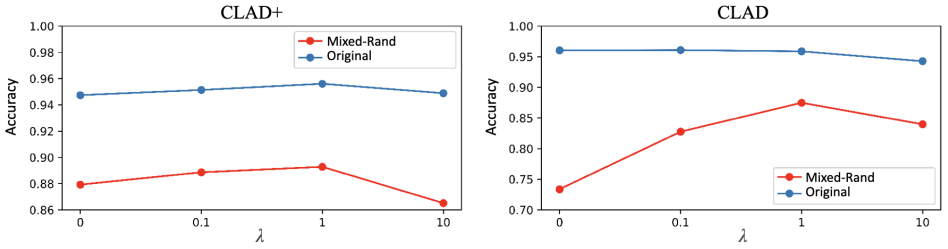
Figure 7: Accuracy on MIXED-RAND dataset and Original IMAGENET-9 with respect to different values of λ

5   Conclusion

Through our work, we present a novel contrastive learning-based approach for background debiasing called CLAD. It samples background-debiased contrastive pairs efficiently. Our work showcases state-of-the-art performance on the Background Challenge dataset. We also show an analysis of our model's features, which explain its superior performance compared to the standard trained model. Further, we empirically demonstrate the need for proper balance between contrastive and supervised losses for the effective debiasing of the model. As a result, training with the proposed contrastive learning method reduces the importance of image background and texture information in the decision-making process of CNN models. Theoretically, this approach works for any discriminative feature pairs, and we took *foreground* vs. *background* and *shape* vs. *texture* (in supplementary material) as an example. In future works, we could further investigate how to extend this approach to other pairs of discriminative features and hopefully guide the CNNs to make decisions based on similar features as humans, thereby improving generalization ability.

References

[1] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 2019.

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. *European conference on computer vision*, 2018.

[3] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. *arXiv preprint arXiv:2110.02797*, 2021.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *International conference on machine learning*, 2020.

[5] Robert Geirhos, David HJ Janssen, Heiko H Schütt, Jonas Rauber, Matthias Bethge, and Felix A Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*, 2017.

[6] 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.

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 2020.

[8] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *International conference on artificial intelligence and statistics*, 2010.

[9] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. *IEEE conference on computer vision and pattern recognition*, 2006.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *IEEE conference on computer vision and pattern recognition*, 2020.

[12] Md Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Björn Ommer, Konstantinos G. Derpanis, and Neil Bruce. Shape or texture: Understanding discriminative features in {cnn}s. *International Conference on Learning Representations*, 2021.

[13] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. *arXiv preprint arXiv:2203.12719*, 2022.

[14] Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. *ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[15] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 2020.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[17] 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.

[18] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social nce: Contrastive learning of socially-aware motion representations. *IEEE International Conference on Computer Vision*, 2021.

[19] Yi-Chen Lo, Chia-Che Chang, Hsuan-Chao Chiu, Yu-Hao Huang, Chia-Ping Chen, Yu-Lin Chang, and Kevin Jou. Clcc: Contrastive learning for color constancy. *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[20] Harshitha Machiraju, Oh-Hyeon Choung, Michael H Herzog, and Pascal Frossard. Empirical advocacy of bio-inspired models for robust image recognition. *arXiv preprint arXiv:2205.09037*, 2022.

[21] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

[22] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. *Advances in Neural Information Processing Systems*, 2021.

[23] Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil Feizi. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[25] Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi, and Pascal Frossard. Hold me tight! influence of discriminative features on deep network boundaries. *Advances in Neural Information Processing Systems*, 2020.

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

[27] 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.

[28] Sepehr Sameni, Simon Jenni, and Paolo Favaro. Dilemma: Self-supervised shape and texture learning with transformers. *arXiv preprint arXiv:2204.04788*, 2022.

[29] Vikash Sehwag, Rajvardhan Oak, Mung Chiang, and Prateek Mittal. Time for a background check! uncovering the impact of background features on deep neural networks. *arXiv preprint arXiv:2006.14077*, 2020.

[30] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *IEEE conference on computer vision and pattern recognition*, 2015.

[32] Saeid A Taghanaki, Kristy Choi, Amir Hosein Khasahmadi, and Anirudh Goyal. Robust representation learning via perceptual similarity metrics. *International Conference on Machine Learning*, 2021.

[33] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018.

[34] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. *IEEE International Conference on Computer Vision*, 2019.

[35] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022.

[36] 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.

[37] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 2019.

[38] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[39] Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 2019.

[40] Zhuotun Zhu, Lingxi Xie, and Alan Yuille. Object recognition with and without objects. *International Joint Conference on Artificial Intelligence*, 2017.