MLVU Final Report
Mitigating Unwanted Background Biases with Background Data Augmentation

Jaeheyoung Jeon , Sooyoung Kim, Jaehwan Lim
Seoul National University
1, Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea
{jan4021,rlatndud0513, amethyst}@snu.ac.kr

Abstract

Based on deep learning, many of the computer vision problems have been solved. However, there are intentional or unintentional biases in the data, and machine learning models trained with these data can also be biased. We note that the background of the image is biased. We have created several datasets by changing backgrounds from existing datasets. We show poor performance when evaluating existing models on these datasets in image classification and object detection tasks. We also finetuned the existing models using these datasets to create a robust model for background changes, and the test performance was higher than the baseline model.

1. Introduction

Since the AlexNet [8] won the first prize of the ImageNet competition in 2012, deep learning has become a de facto standard for image classification and object detection models. Deep neural networks are achieving state-of-the-art results on various applications in computer vision problems. However, due to the nature of the black-box model, people do not know why their model made such predictions, so they do not know whether they trained the model correctly as they intended or just trained to fit benchmark datasets. And deep learning based models are easily fooled by small changes in input images [13]. Deploying machine learning models in the real world requires models to be robust to change, especially in safety-critical applications.

'Garbage In Garbage Out' means that poor data quality produces unreliable models and results. In a machine learning model based on a lot of data, the quality of the data is a very important issue. In particular, the bias problem of training datasets has been receiving attention continuously, and overcoming it is a challenging problem. It is well known that deep learning models trained with biased data yield biased outcomes, and we call this 'Bias In Bias Out'.

There are several unwanted biases in the dataset that are frequently used for training the image classification model, and we try to solve the background bias problem. In [14], authors trained a machine learning model that distinguishing between wolves and huskies. They intentionally hand selected such that all pictures of wolves had snow in the background, while pictures of huskies did not. The model predicts wolf if picture is in a snowy background, and husky otherwise. Because the model has learned its background, it may be necessary to build an datasets to prevent training such a bad model. However it takes a long time and a lot of cost to newly build an dataset, we propose to overcome this problem through data augmentation.

We construct a dataset with a different background only from the existing image benchmark dataset and evaluate the image classification and object detection model there. We apply image segmentation models to images to segment objects from background. Here we used pre-trained image segmentation model [11]. Then we apply three different background augmentation methods. The first method is to change the background to the background of another image. The second method is to change the background to a single color, red, green, blue and black. The third method is to change the background to average of the background of the data set.

We will show existing deep learning based image classification and object detection models are biased towards background. Further more, we will show the image classification and object detection model will show robustness to background changes, when training is performed on our augmented dataset.

2. Related Work

In [6], researchers create datasets for two other forms of robustness, ImageNet-C dataset for input corruption robustness and the ImageNet-P dataset for input perturbation robustness. ImageNet-C datasets consists of algoritmically

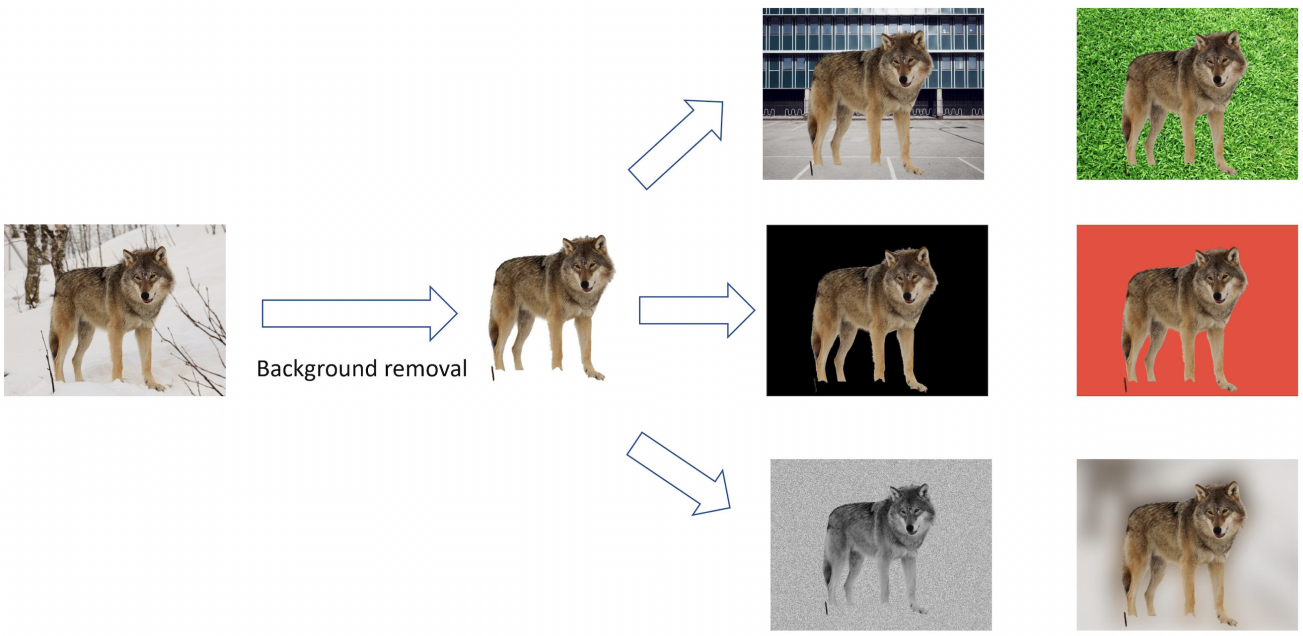

Figure 1. Proposed background data augmentation method. Image segmentation models are used to distinguish backgrounds from objects. For datasets that already have semantic segmentation annotation, we omit this process. Then we apply three background augmentation methods to images.

generated corruptions from noise, blur, weather and digital categories. ImageNet-P datasets are generated by having pertubation sequences from each ImageNet validation image. Image classification models showed significant performance degradation on simple perturbations and authors introduced methods that improve robustness to perturbations.

In [4], authors show that ImageNet-trained Convolutional Neural Networks are strongly biased towards recognising textures rather then shapes. They create a dataset by striping image from ImageNet of its texture and replacing it with the style of a randomly selected painting through style transfer. With this dataset, authors did some experiments and concluded that the texture bias in current CNNs was not due to the structural problem of CNNs, but due to the training data of the ImageNet.

Data augmentation is common technique of increasing the amount and diversity of data set for training a classifier. In the image domain rotation, random cropping, image mirroring, color jittering, adding a noise and using elastic distortions are common data augmentation policy. In [2], authors use Reinforcement Learning as the search alorithm to automate the process of finding an effective data augmentation.

Recent study [15] studied techniques to apply group distributionally robust opimization problem to over-parametrized neural networks. Machine learning models can learn spurious relations, such as the background of images in test data and relationships with answer labels, which significantly degrades the accuracy of machine learning models in the real world. To prevent machine learning models from learning false relationships between background and answer labels, authors in [15] constructed a new dataset, Waterbirds, which changes the background from bird photographs on the Caltech-UCSD Birds-200-201 (CUB) dataset [17] to background image from the Places dataset [18]. They argued that regularization is important for training models that generalize from worst-case test datasets.

In [12], researchers studied to reduce background bias through background substitution in person re-identification problem. They argued that dataset for re-identification is small then model trained on that dataset is greatly affected by the background of the images. To segment the person from the background, they used a deformable parts model or background substraction technique. Then they applied a simulated background to segmented person images. In more recent studies [16], researchers studied background-bias problem in anomaly detection in surveillance videos. They questioned that whether deep learning models really learn the anomaly situation or just remember the background-bias.

Figure 2. Example images from CIFAR-10

Figure 3. Example images from Microsoft COCO

3. Datasets

For image classification task we developed a proof of concept by using CIFAR-10 [7]. The CIFAR-10 are labelled subsets of the 80 million tiny images dataset. CIFAR-10 is commonly used for image classification task. The CIFAR-10 dataset consists of 60000 32x32 color images, with 6000 iamges per class and there are 50000 training images and 10000 test images. There are 10 completely mutually exclusive classes.

We used Microsoft COCO 2017 [9] and PASCAL VOC 2012 challenge datasets[3], changed the background. COCO is richly-annotated dataset comprised of images depicting complex everyday scenes of common objects in their natural context. It focus on segmenting individual object instances. It contains 91 common object categories, in total the dataset has 2,500,000 labeled instances in 328,000 images. The PASCAL VOC 2012 is a very popular dataset for evaluating algorithm for object detection. It contains 20 object categories, in total the dataset has 27450 annotated objects in 11540 images. Since MS-COCO and PASCAL VOC have already object segmentation annotation, we will omit the image segmentation process.

4. Methods

4.1. Overview

The two questions that we want to investigate are 1) Existing machine learning models are biased towards background? 2) If so, how can we make a model that is robust to changes in the background without being biased? CNNs learns the pattern of input images and the background affects to the image classification, we could construct a new dataset by changing the background so that there is no specific pattern. We will evaluate the performance of the existing models on our dataset.

4.2. Image segmentation

We first used FCN-ResNet101 image segmentation model to distinguish objects from backgrounds. FCN-ResNet101 is constructed by a Fully-Convolutional Network model with a ResNet-101 backbone [5]. In [11], authors transform all fully connected layers into convolution layers and append 1 x 1 convolution with channel dimensions 21 to predict scores. And they introduce a skip architecture that efficiently combines high-level semantic information and local appearance information. However, segmentation performance was not so good on CIFAR-10, so we decided to use a different model, DeepLabv3-ResNet101[1]. Deeplabv3-ResNet101 is constructed by a Deeplabv3 model with a ResNet-101 backbone.

If there is already segmentation annotation per objects, like Microsoft COCO, we skip this process.

4.3. Background augmentation

Currently, we do not know which background augmentation method can effectively reduce the background bias. We design three background augmentation methods. We

will change the background of the target image to randomly selected the background of other images, a single color, or black and white noise background.

5. Experiment

5.1. Image classification

5.1.1 Overview

We experiment with test dataset to verify out hypothesis. And to resolve this, in training step we augment data with three criterions, background, threshold and transfer learning.

Training data is augmented data by our methods and 50000 CIFAR-10 train data is added whether pre-trained or not. Test data is always CIFAR-10 test dataset. More details are in 5.1.3 Training.

5.1.2 Hypothesis Verification

We first experiment with changing background images from CIFAR-10 to a single color.

In CIFAR-10, images are of size 32 x 32 x 3, we resize images to 224 x 224 x 3 and normalize it with mean = 0.5 and std = 0.5. Then we apply Deeplabv3-ResNet101 image segmentation model to the normalized images. And then resize the output to 32 x 32. Using these semantic labels we change the background to a single color.

ResNet : In Table 1, we experiment with the effect of changing image background to a single color. As shown in the table, in the case of Original, the test accuracy is 88.49%. We sorted 5643 images from CIFAR-10 test data with threshold 150 and evaluated the ResNet on the red, green, blue and black background with that 5643 images, performance drops from 88.49% to 53.85%, 72.37%, 66.01% and 63.06%, respectively.

We show that ResNet underperformed on our datasets. We will change the background in other ways and measure the performance of other popular models.

Background	Original	Red	Green	Blue	Black
Test Acc (%)	88.49	53.85	72.37	66.01	63.06

Table 1: Test Results.

5.1.3 Training

In training, we start with set our three criterions. First, in background, we augment our data as changing background pixels into red, green, blue and black 4 color backgrounds or 7 backgrounds to represent each classes. The seven backgrounds are images of sky, sea, branch, floor, forest, meadow and road. They are collected to represent each class in CIFAR-10. But some classes like dog do not have any backgrounds to be considered as representative. And dog's backgrounds overlap with other classes' background and are very various as floor at home, meadow, forest, and so on. So we choose these classes' background as floor. Second, in threshold, with segmented train dataset of CIFAR-10 we choose only images whose number of sorted pixels as objects are more than 150. We also select the images by human labor extracting well segmented 1015 images after checking each of the first 5000 images among the segmented train data with our eyes. Sorting whole 50000 CIFAR-10 train data takes too much time. 5000 images are classified as best as possible in time. Third, in transfer learning we just choose to transfer learning or start training at the bottom.

Overall, training at the bottom is better than transfer learning and human threshold is better than 150 threshold. Two cases out perform the origin pre-trained ResNet with 88.49% First, the model trained from scratch with human threshold and 4colors outperforms as 88.77%. We here use 50000 + 1015x4 data to train ResNet. Second, the model trained from scratch with human threshold and 7 backgrounds outperform as 89.04%.

Background	Original	RGB+Black	7Background
Test Acc(%)	88.49	88.77	89.04

Table 2: Train Results : This results obtained from human threshold and training from scratch.

5.1.4 Conclusion

From experiments, we have proved that background has impact on the accuracy of model. And with data augmentation changing background, we outperform origin models for several times. In Image Classification, both RGB+Black and 7 Backgrounds outperform origin ResNet. Even we use only one-fiftieth of total 1015 images of 50000 train data, we get these outperforming result. The time limit allowed us to use only 1015 images to augment, but we strongly expect better result with fully well segmented 50000 data.

5.2. Object detection

5.2.1 Overview

We train and evaluate our models with PASCAL VOC 2012 having 20 classes. As only 2913 images have annotations among PASCAL VOC 2012, we pre-train the faster-rcnn model using MS COCO datests and fine-tune the model our data. To verify the impact of background in object detection, PASCAL VOC and COCO with background changes test a few models including Faster RCNN.

Figure 4. Sample images from CIFAR-10 with various backgrounds: Red,Green,Blue,Sky,Branch and Road.

Figure 5. Sample images from PASCAL VOC with human select backgrounds.

Various backgrounds and two test datasets which are original dataset and augmented dataset are criterions to augment data and train/test models. We also combine datasets having different backgrounds which get the highest mAP scores.

5.2.2 Hypothesis Verification

We first experiment object detection models on PASCAL and COCO. In PASCAL VOC 2007 test data, 210 images are annotated with segmentation information. We changed the background to red, green, blue and black for 210 images, respectively. We then evaluated Faster R-CNN and RetinaNet[10] on these datasets with IoU threshold = 0.5. And we also changed the background 5000 images COCO 2017 validation dataset, and evaluated Faster R-CNN, YOLO-v3 and SSD on augmented datasets. We denoted AP, AP_{50}, AP_{75} by average precision at IoU = 0.50:0.05:0.95, 0.50, 0.75 respectively. We used the Faster R-CNN , RetinaNet, YOLO-v3, and SSD implementation from https://github.com/open-mmlab/mmdetection with backbone models are ResNet-101, Resnet-50, DarkNet-53 and VGG16 respectively. All hyperparameters kept at default(Table 3 and 4). All the tested models performed poorly on images with augmented backgrounds.

5.2.3 Training

We train models with not augmented original PASCAL VOC 2012 data. Considering the number of augmented datasets, datasets increased the amount of original data by 2, 5, 6, and 19 times are used as baselines. Dataset is doubled in amount for comparison with black and mean value backgrounds. Datasets increased by 5 and 6 times in amount are used for comparison with rgb+black background data and rgb+black+mean value background data. Data increased by 19 times in amount is compared with data augmented with human select backgrounds. The performances usually increase as more data is used (Table 5).

Test dataset

We test our background models in two dataset. In all kinds of background models, although the model is trained with original and augmented data, it detects objects better in testing using augmented data than original data. It is due that the number of the augmented data is much bigger than the number of original data in RGB+Black background model and human select background model (Table 6). In black background model and mean value background models, the models in which the numbers of original data and augmented data are the same, detecting objects with new data having simplified backgrounds is easier than data having various and not consistnet backgrounds. Therefore we focus on augmented data testing results (Table 6 and 7).

Background

We tried three methods to augment the background of the data. First, we simplify the background by applying

Model	Metric	Original	Red	Green	Blue	Black
Faster RCNN	AP	0.374	0.324	0.318	0.315	0.324
	AP_{50}	0.581	0.475	0.466	0.466	0.475
	AP_{75}	0.404	0.355	0.349	0.349	0.371
YOLO-v3	AP	0.337	0.294	0.286	0.288	0.313
	AP_{50}	0.566	0.448	0.431	0.437	0.497
	AP_{75}	0.353	0.319	0.312	0.313	0.319
SSD	AP	0.294	0.275	0.265	0.265	0.287
	AP_{50}	0.493	0.418	0.401	0.401	0.447
	AP_{75}	0.310	0.300	0.287	0.289	0.312

Table 3: The mAP scores on COCO 2017 validation set

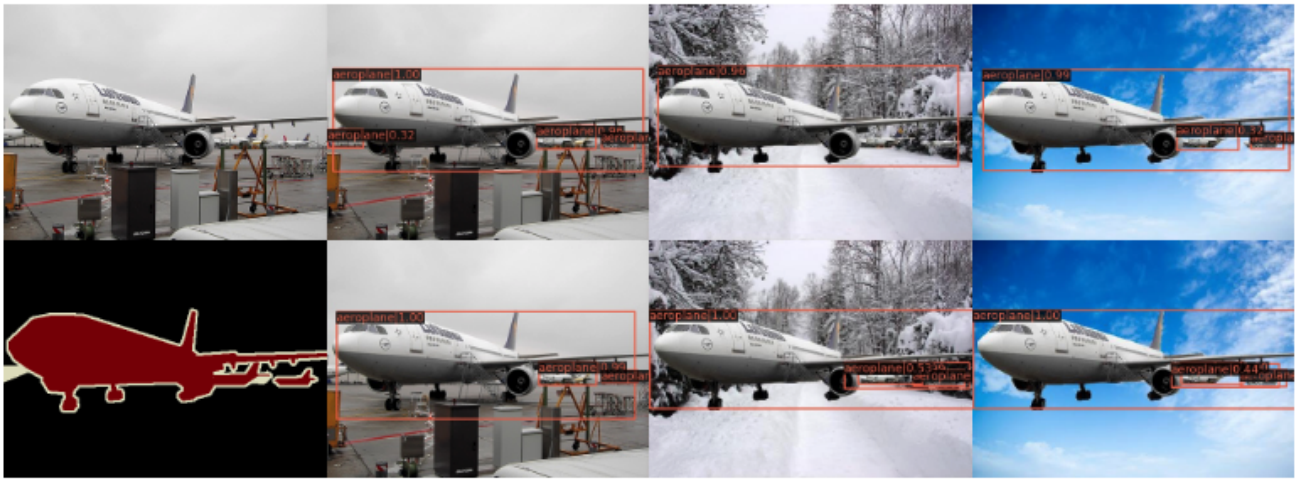

Figure 6. Qualitative analysis. Top: Original image, evaluate baseline model on original image, snow background and sky background respectively. Bottom: Segmentation image(Ground truth, evaluate our model(fine-tuned using human select background) on original image, snow background and sky background respectively.

	Original	Red	Green	Blue	Black
Faster R-CNN	0.804	0.715	0.675	0.695	0.700
RetinaNet	0.769	0.692	0.633	0.684	0.696

Table 4: The mAP scores on PASCAL VOC 2007 test set

	2 times	5 times	6 times	19 times
# of images	4370	10925	13110	41515
mAP	0.707	0.762	0.735	0.767

Table 5: The mAP scores of baselines trained with original PASCAL VOC 2012 data by increasing the amount.

solid colors including black, red, green and blue. When we use black background which is the easiest method, the mAP outperforms the baseline. Then we change the background color with red, green, blue and black and combine them into train data. The mAP performance of applying RGB+Black background gets a higher score than not only baseline (5 times) but also black background in testing only using augmented data.

Second, we reflect diversity of backgrounds by selecting 13 backgrounds containing sky, snow, swimming pool, sunset, city, auditorium etc which objects in PASCAL VOC 2012 data usually appear at. We propose that as objects appear mainly in the background where they are usually located, backgrounds and objects have strong associations. So, various backgrounds we chose can reflect different backgrounds in original PASCAL data. Using human select backgrounds gets higher mAP performance than baseline in testing only on augmented data.

Third, we integrate the above two methods to simplify

	Test dataset	Black	RGB + Black	Human select background	Mean	Mean without original
Baseline	Original	0.707	0.762	0.767	0.707	0.707
Augmentation model	Original	0.699	0.757	0.688	0.767	0.619
	Augmented	0.787	0.845	0.789	0.828	0.819

Table 6: The mAP scores of the data augmented with various backgrounds.

	Test dataset	RGB+Black+Mean	RGB+Black+Mean without original
Baseline	Original	0.735	0.735
Augmentation model	Original	0.733	0.56
	Augmented	0.820	0.834

Table 7: The mAP scores of the combined data augmented with various backgrounds.

and reflect various backgrounds in PASCAL data. The results of two methods mean that both simplification and diversity of background affect object detection. To simplify various backgrounds of all data, we create a mean value background by calculating the average values of all pixels in backgrounds. The mean value background has the highest and most improved mAP results in testing only on augmented data.

Among training data augmented by using mean value background, original PASCAL data is the only dataset not having a simplified background. To find the effect of mean value background, we exclude original PASCAL data and use only augmented data. Different backgrounds of original data interfere with training the effect of the simplified various backgrounds to object detection we want to test. Therefore we focus on augmented data testing results. When we train the model using a dataset augmented with the mean value of each pixel, the mAP score is the second highest result after the mean value model including original PASCAL data in testing only using augmented data. We find that not only simplifying various backgrounds but also a variety of original backgrounds is important to detect objects.

Among the models trained and tested in three methods, the two models having the highest results are RGB+Black and mean value background models. We combine two datasets in each model to increase diversity of the backgrounds. We train in two cases and test with two data. In training, we build two models including or excluding original PASCAL data to include or exclude original data having backgrounds that are not consistent with augmented data. In testing, original data and augmented data are used. In all cases, a model trained using RGB+Black and mean value background data without original PASCAL data gets the highest score in testing only with augmented data. It means that solid colors and mean value simplified various backgrounds help the model to learn how to detect objects well and detect objects better in new images than any other models.

5.2.4 Conclusion

We propose a method simplifying background by using solid colors, a method reflecting various backgrounds, and a method simplifying various backgrounds to ignore the effect of background to object detection. The models using data augmented based on background showed mAP performances range from 0.56 to 0.834. Among many models, the models simplifying the background reflecting various backgrounds or not have the highest mAP score and the most improved results than baselines. We find that background affects object detection and simplifying, not diversifying background helps the model to detect objects.

6. Discussion

In Classification, although we apply one of the well used models for segmentation, the results from segmentation are not accurate. When the state of the art model in segmentation is developed, the future study can try classification again with background augmentation. In object detection, training data only has 2,913 data having annotation among 17,125 data and testing data has no annotation data in PASCAL VOC 2012 dataset. Therefore, when the number of segmented data with annotation increases, future study can get precise results.

7. Supplementary Material

In here we use only well segmented 1015 data which is sorted by human labor to train ResNet. And test data 10000 CIFAR-10 test data same here. With 1015 images ResNet reaches 45.93% accuracy. With augmented by 4color, 1015x(4+1) data makes ResNet reach to 49.57%. Similarly, with 7 backgrounds and 1015x(7+1) data ResNet has reached 50.33%

Model	Human + Mean	Total(Original + RGB + Black + Human +Mean)
Baseline	0.767	0.767
Augmentation model	0.594	0.710

Table 8: The mAP results of combined models tested only with original data

	Test dataset	Black without original	RGB+Black without original
Baseline	Original	0.707	0.707
Augmentation model	Original	0.384	0.328
	Augmented	0.779	0.794

Table 9: The mAP results of solid background models trained only with augmented data without original data

Background	Original	RGB+Black	7Background
Train data	1015	1015x(1+4)	1015x(7+1)
Test Acc (%)	45.93	49.57	50.33

Table 10: Experiment results

Among combined models, models using human select background get lower results than baselines(Table 8). Solid background models have mAP scores range from 0.328 to 0.794. Testing with augmented data gets higher performances than baseline models(Table 10).

References

[1] L. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[2] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501, 2018.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[4] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2019.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[6] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019.

[7] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[9] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection, 2018.

[11] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.

[12] N. McLaughlin, J. M. Del Rincon, and P. Miller. Data-augmentation for reducing dataset bias in person re-identification. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2015.

[13] A. M. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897, 2014.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[15] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731, 2019.

[16] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang. Eliminating background-bias for robust person re-identification. pages 5794–5803, 06 2018.

[17] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[18] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.