Supplementary Material for Partition-and-Debias: Agnostic Biases Mitigation via A Mixture of Biases-Specific Experts

Jiaxuan Li
The University of Tokyo, Japan
li@nlab.ci.i.u-tokyo.ac.jp

Duc Minh Vo
The University of Tokyo, Japan
vmduc@nlab.ci.i.u-tokyo.ac.jp

Hideki Nakayama
The University of Tokyo, Japan
nakayama@ci.i.u-tokyo.ac.jp

This supplementary material complements our paper with the following parts: First, we present details in implementing our PnD (Sec. A), which are not included in the main paper. Second, we provide more details on the dataset used in the main paper and some concepts related to biases (Sec. B). Finally, we add more analysis to assess PnD (Sec. C).

A. Implementation Details

Network structure. The main network structure is shown in Tab. A, including details in layer names, input sizes and operations. Experts 1 – 4 are the inserted biases-specific experts in PnD, the remainings denote the layers and blocks of the debiased/bias encoder in it, which is the same with ResNet-18 [2]. In the third operation of each expert, the linear layer input dimension is the doubled dimension from adaptive avgpool, since it will take the concatenation of debiased features and bias features as the input. The BatchNorm and ReLU operations are not shown in this table for simplicity. Only one kind of debiased/bias encoder is given in the table because their structures are the same.

Diversity loss. The KL $\left(\hat{\mathbf{y}}_{\mathrm{b}}^{(i)}, \hat{\mathbf{y}}_{\mathrm{b}}^{(i-1)}\right)$ in diversity loss $\mathcal{L}_{\text{div}}$ is calculated as:

$$\mathrm{KL}\left(\hat{\mathbf{y}}_{\mathrm{b}}^{(i)}, \hat{\mathbf{y}}_{\mathrm{b}}^{(i-1)}\right) = \hat{\mathbf{y}}_{\mathrm{b}}^{(i)} \log\left(\frac{\hat{\mathbf{y}}_{\mathrm{b}}^{(i)}}{\hat{\mathbf{y}}_{\mathrm{b}}^{(i-1)}}\right),$$

where $\hat{\mathbf{y}}_{\mathrm{b}}^{(i)}$ denotes the bias prediction distribution from i^{th} expert, $\hat{\mathbf{y}}_{\mathrm{b}}^{(i-1)}$ denotes that from the previous one.

Training details. In all experiments, we train our framework with two-phase optimization using Adam with a batch size of 128. Following settings in [8], the hyperparameter q in GCE loss is set to 0.7.

Figure A. Examples from CelebA dataset.

For {Biased MNIST, BAR, Modified IMDB, MIMIC-CXR + NIH}, during the initial and counterfactual training stages, we set the epoch number as {70, 70, 20, 10} and {100, 70, 30, 50}, α as {0.2, 0.6, 0.2, 0.2} and {2.0, 1.0, 2.0, 2.0}, learning rate (LR) as {1e-3, 1e-4, 5e-4, 5e-4} and {5e-4, 5e-5, 5e-4, 1e-4}, LR is decent every {20, -, -, -}, and {20, 10, -, 30} epochs with an LR Decay Gamma of 0.5, respectively. In the second stage, β is set as {4.0, 0.3, 0.3, 0.3}, K is 16, P is 8, and the temperature hyperparameter in contrastive loss is {0.1, 0.07, 0.1, 0.1}. In both two stages, weight decay is {1e-5, 5e-6, 1e-6, 1e-6}. We do not change any hyperparameters when the bias ratio is different for a certain dataset. "-" means no specific value.

For additional experiments on CelebA, during the initial and counterfactual training stages of PnD , we set the epoch number as 10 and 20, α as 0.2 and 2.0, respectively. In both two stages, learning rate (LR) is 5e-4, weight decay is 1e-6. In the second stage, β is set as 0.3, K is 16, P is 8, LR is decent every 10 epochs with an LR Decay Gamma of 0.5.

B. Datasets

B.1. Public datasets

CelebA [7]. It is a publicly available face attribute dataset that contains 202599 face images of 10177 celebrity identi-

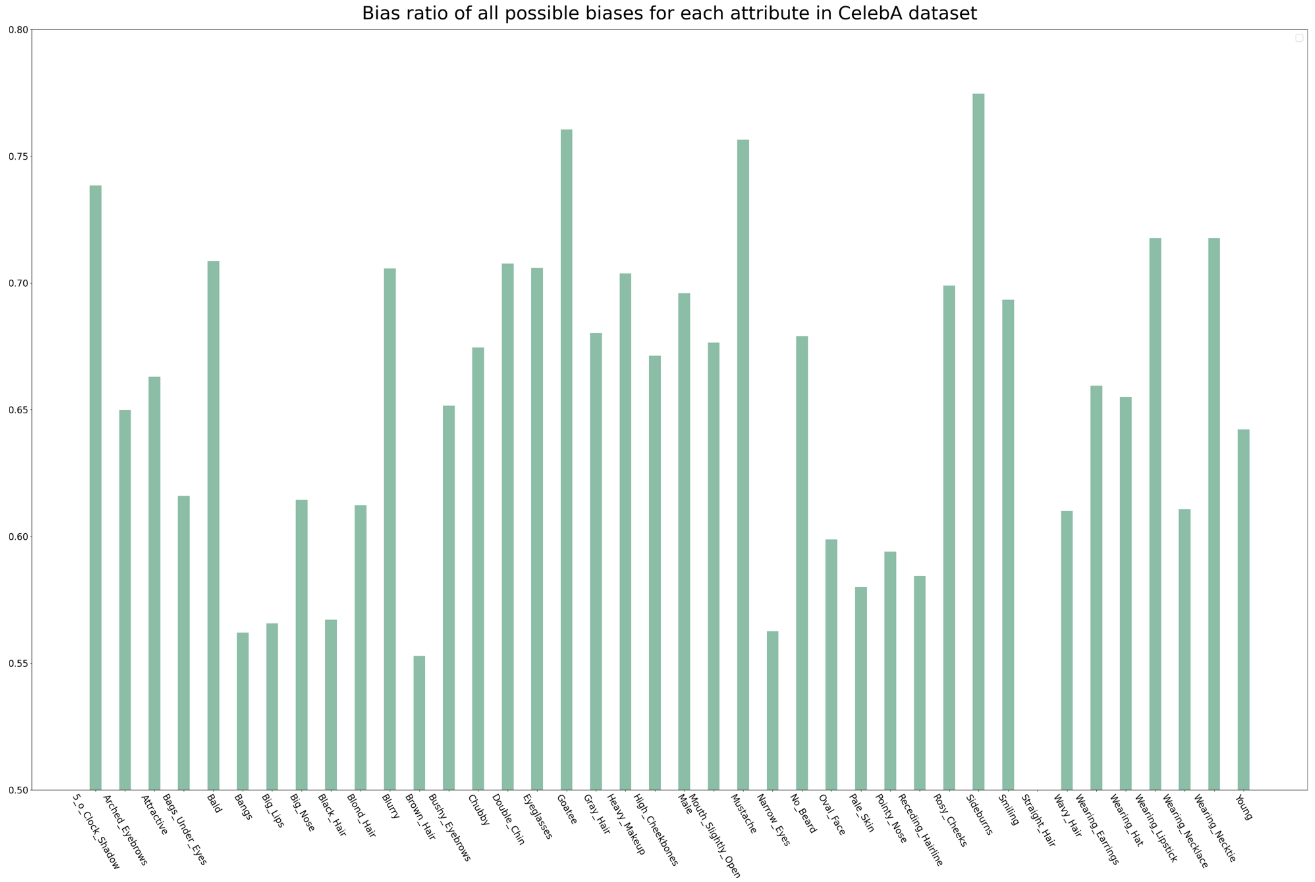

Figure B. Bias ratio of all possible biases for each attribute in CelebA dataset. We can find that most attributes are biased. Note that in this figure (also in other figures), we use the original annotation from CeleA as the name of attribute.

ties, each with 40 attribute annotations (Fig. A).

Take the age attribute for example, if most images with *young* category are annotated with *female*, while most images with *old* category are annotated with *male*. We can consider the gender attribute as a bias for age attribute. In this bias scenario, we have two concepts relevant to the bias problem:

(i) Bias ratio. It denotes the probability of co-occurrence of the bias category and the target category. For *young* in age attribute, the bias ratio of *female* in gender bias attribute refers to the proportion of individuals with both *female* and *young* in the total number of individuals with *young*.

(ii) Worst group. Following the definition in [6], the worst group in this paper denotes the group that gets the lowest accuracy score among all 4 combinations of the target categories and the bias categories, such as (*young*, *old*) × (*male*, *female*).

As shown in Fig. B, we analyze this dataset by calculating the bias ratio of possible biases for each attribute in this dataset. That is, for each target attribute, we analyze the percentage of other attributes within each of its categories (here, two categories), and if certain other attribute would be a bias as described above, we calculate the bias ratio for each category of the target attribute as in Fig. C. For each target attribute, all bias ratios of possible biases are averaged. From Fig. B, we can see most attributes are biased in CelebA.

Biased MNIST [10]. It contains 10 digits (*0* – *9*) as its target categories and 7 biases: digit color, digit scale, digit position, type of background texture, background texture color, co-occurring letter, and letter color (Fig. D). There are 50000, 10000, and 10000 images for training, validation, and test.

BAR [8]. There are typical action-place pairs, including *climbing* and *rockwall*, *fishing* and *water surface*, *diving* and *underwater*, *vaulting* and *sky*, *racing* and *a paved track*, *throwing* and *playing field* in the 1941 training images (Fig. E, 1st and 2nd cols); and unseen samples beyond the settled pairs in the 654 test images (Fig. E, 3rd col). BAR can be seen as a dataset with a single bias, where the action is spuriously correlated with the background.

Table A. The main network structure of PnD. We omit BatchNorm and ReLU operations in this table. Although PnD has four debiased blocks, four bias blocks, and four biases-specific experts with a debiased classifier and a bias classifier. We only show the debiased/bias parts here since the two parts have the same structure.

Layer name	Intput size	Operation
Initial layer	$160 \times 160 \times 3$	$[7 \times 7, 64] \times 1$ conv
Initial pooling	$80 \times 80 \times 64$	$[3 \times 3]$ maxpool
Block 1	$40 \times 40 \times 64$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$ conv
Expert 1	$40 \times 40 \times 64$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 512 \end{bmatrix} \times 1$ conv
		adaptive avgpool
		$\begin{bmatrix} 1024, 16 \\ 16, 10 \end{bmatrix} \times 1$ linear
Block 2	$40 \times 40 \times 64$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$ conv
Expert 2	$20 \times 20 \times 128$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 512 \end{bmatrix} \times 1$ conv
		adaptive avgpool
		$\begin{bmatrix} 1024, 16 \\ 16, 10 \end{bmatrix} \times 1$ linear
Block 3	$20 \times 20 \times 128$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$ conv
Expert 3	$10 \times 10 \times 256$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 512 \end{bmatrix} \times 1$ conv
		adaptive avgpool
		$\begin{bmatrix} 1024, 16 \\ 16, 10 \end{bmatrix} \times 1$ linear
Block 4	$10 \times 10 \times 256$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$ conv
Expert 4	$5 \times 5 \times 512$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 512 \end{bmatrix} \times 1$ conv
		adaptive avgpool
		$\begin{bmatrix} 1024, 16 \\ 16, 10 \end{bmatrix} \times 1$ linear

B.2. Our contructed datasets

Modified IMDB. Original IMDB [9] contains 460723 face images from 20284 celebrities. We use the cleaned IMDB dataset with 112340 face images provided by [4], which is filtered by pretrained models designed for age and gender classification. Their cleaned IMDB consists of extreme bias 1 (EB1), extreme bias 2 (EB2), and a test set. EB1 contains 36004 face images, which are old people (aged 40+) and men, or are young people (aged 0 – 29) and women. EB2 contains 16800 face images, which are opposite to EB1. The test set contains 13129 images, all are aged 0 – 29 or 40+ without other settings. The gender attribute is a bias for the age classification in the cleaned IMDB dataset. In

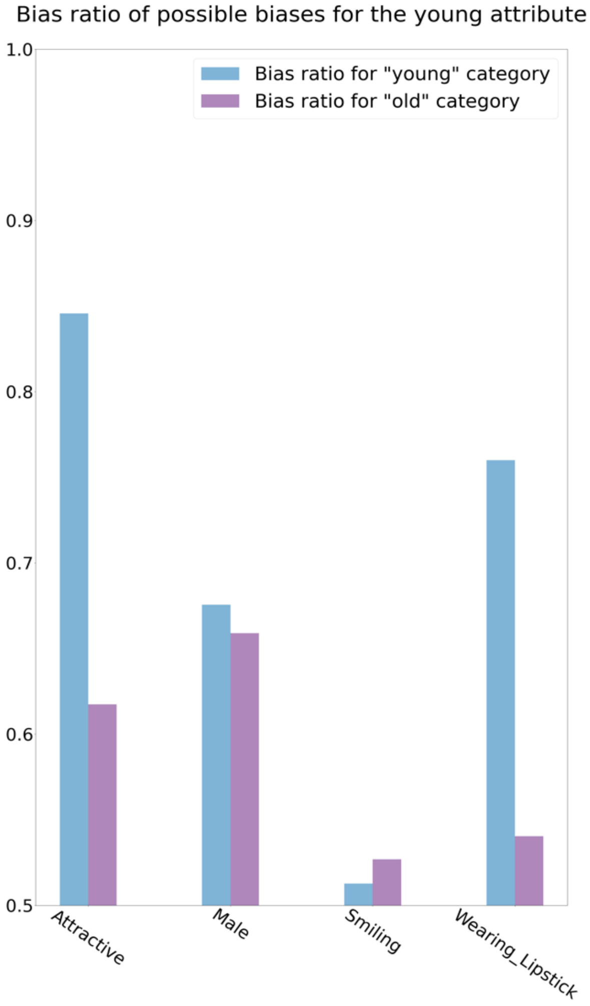

Figure C. Bias ratio of all possible biases for the age attribute in CelebA.

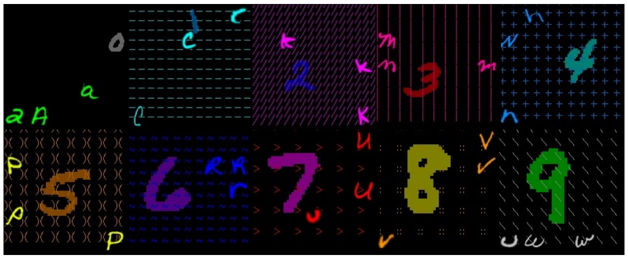

Figure D. Examples for digit *0* – *9* from Biased MNIST training data, there are 7 biases in this dataset, including digit color, digit scale, digit position, type of background texture, background texture color, co-occurring letter, and letter color.

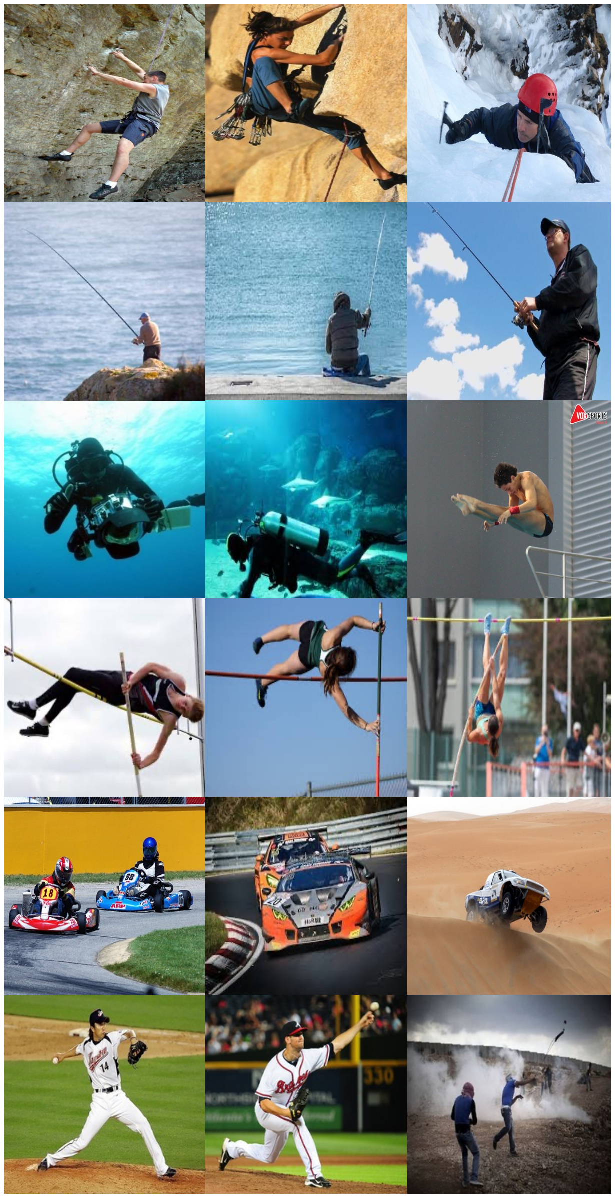

Figure E. Examples from BAR dataset. The images in 1st and 2nd cols are training data with action and corresponding co-occurring background pairs, including climbing and rockwall, fishing and water surface, diving and underwater, vaulting and sky, racing and a paved track, throwing and playing field (from 1st – 6th rows). The images in 3rd col are test data, all of which are scenarios not seen in the training set.

order to add another natural bias to this dataset, so that it has two biases for age classification. As shown in Fig. F, we use a face attribute transfer model [1] to put glasses on the faces. Thus, we can make *wearing glasses* also a bias in this dataset by controlling the ratio of *wearing glasses*. As a result, we create Modified IMDB, where most images *young* labels are *female* and *wearing glasses* and most images with *old* labels are *male* and *not wearing glasses*. It consists of 20000 training images with a bias ratio of 0.95 and 0.99, 1617 unbiased validation images, and 1617 unbiased test images.

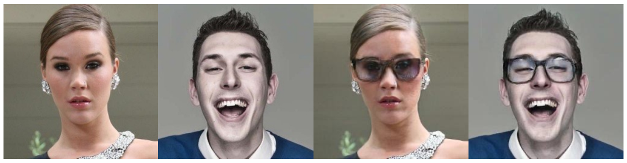

Figure F. The orginal images (left two) from IMDB and the images with glasses (right two) generated by a face attribute transfer model.

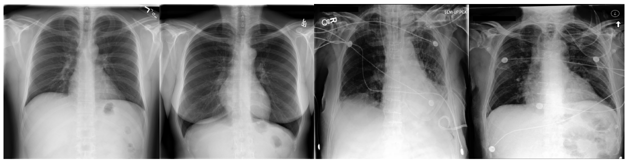

Figure G. Examples from MIMIC-CXR + NIH dataset. They are source-biased, the left two are from NIH with *no finding* labels, and the right two are from MIMIC-CXR with *pneumonia* labels.

MIMIC-CXR + NIH. It is constructed by simulating biases brought by different data sources when collecting datasets. We mix MIMIC-CXR [3] and NIH [11] datasets into a MIMIC-CXR + NIH dataset. The original NIH contains 50500 *no finding* and 876 *pneumonia* training images, 9861 *no finding* and 555 *pneumonia* test images. The original MIMIC-CXR has 10145 *no finding* and 7209 *pneumonia* training images, 122 *no finding* and 140 *pneumonia* test images. Considering *pneumonia* images are very few in NIH, we construct MIMIC-CXR + NIH by collecting most *pneumonia* images from MIMIC-CXR, while most *no finding* images from NIH (Fig G). In MIMIC-CXR + NIH, the target categories are *no finding* and *pneumonia*, and the biases come from two data sources. It contains 8500 training images with a bias ratio of 0.80 and 0.95, 500 unbiased validation images, and 500 unbiased test images.

Multiple Biased MNISTs. This set of datasets is created according to the method in constructing Biased MNIST [10]. It consists of 7 Biased MNIST datasets with different numbers (ranging from 1 to 7) of biases. As shown in Fig. H, we construct this set of multiple Biased MNISTs by gradually adding digit color, digit scale, digit position, texture, texture color, letter, and letter color biases (from 1st – 7th rows) into MNIST [5].

C. Additional Detailed Analysis

Visualization for learned target and bias features. In Fig I, we visualize the region of interest on more examples from the Modified IMDB dataset. The upper one is

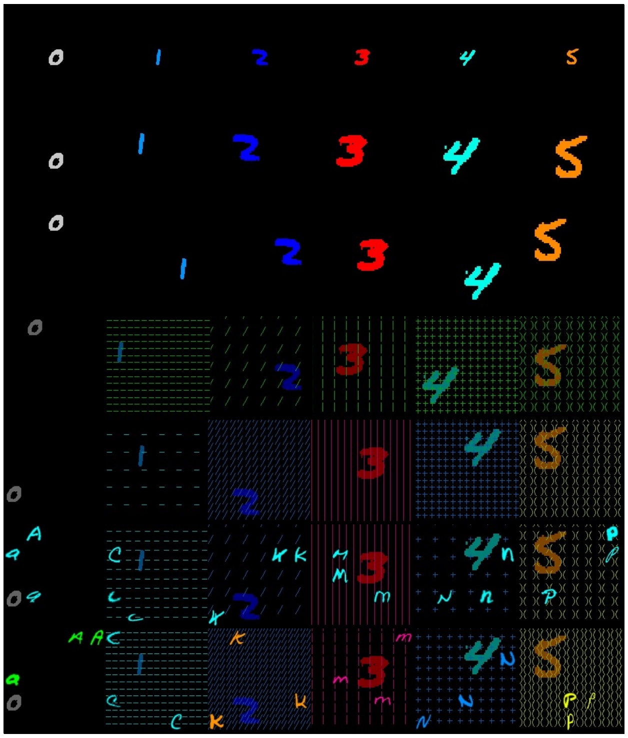

Figure H. Examples from multiple Biased MNISTs with a different number of biases, 1st – 7th rows indicate the number from 1 – 7 (gradually adding digit color, digit scale, digit position, texture, texture color, letter, and letter color biases). Note that we only show the same samples with 0 – 5 digits here in all cases for clarity, we have 10 digits in total.

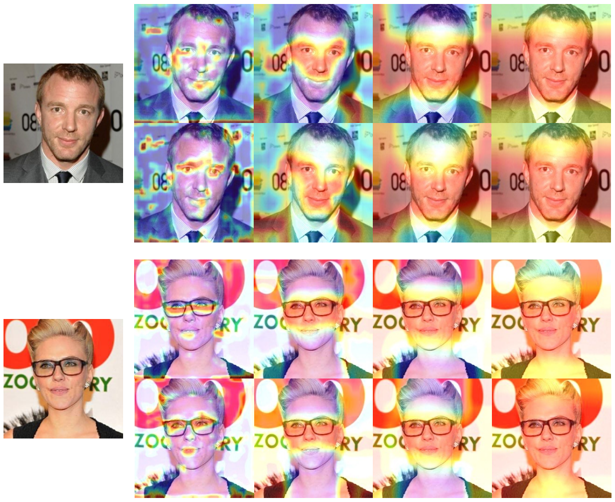

Figure I. Regions of interest for biases-specific experts of our PnD in the debiased (1st and 3rd rows) and bias (2nd and 4th rows) encoders, when conducting action classification in the test set of Modified IMDB. The original images are in the 1st column, and 2nd – 5th columns are their saliency maps generated using Grad-CAM, from the first expert to the fourth expert. The regions of interest for debiased classification and bias detection are changing as the network gets deeper, and there are also significant differences between the two tasks.

an image of a young male without glasses, which conflicts with the bias samples (*young*, *female*, *wearing glasses*) in the training set. The lower one is aligned with bias samples. We can see both debiased age classification and bias detection focus on varying level features in different depths. At the same time, there are differences between these two parts. In the first expert's results, debiased age classification is more related to the position under the eyes, which may be related to age. In contrast, bis detection concentrates on the glasses region.

References

[1] Xuyang Guo, Meina Kan, Zhenliang He, Xingguang Song, and Shiguang Shan. Image style disentangling for instance-level facial attribute transfer. *Computer Vision and Image Understanding*, 207:103205, 2021. 4

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[3] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019. 4

[4] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[5] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 4

[6] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *Proc. European Conference on Computer Vision (ECCV)*, 2022. 2

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[8] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[9] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018. 3

[10] Robik Shrestha, Kushal Kafle, and Christopher Kanan. Occamnets: Mitigating dataset bias by favoring simpler hypotheses. In *Proc. European Conference on Computer Vision (ECCV)*, 2022. 2, 4

[11] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4