PRIVACY-PRESERVING OBJECT DETECTION

Peiyang He[1,*] **Charlie Griffin,**[1,*] **Krzysztof Kacprzyk,**[1,*] **Artjom Joosen,**[1]

Michael Collyer,[1] **Aleksandar Shtedritski,**[1] **Yuki M. Asano**[1 †]

[1] Oxford Artificial Intelligence Society, University of Oxford

ABSTRACT

Privacy considerations and bias in datasets are quickly becoming high-priority issues that the computer vision community needs to face. So far, little attention has been given to practical solutions that do not involve collection of new datasets. In this work, we show that for object detection on COCO, both anonymizing the dataset by blurring faces, as well as swapping faces in a balanced manner along the gender and skin tone dimension, can retain object detection performances while preserving privacy and partially balancing bias.

1 INTRODUCTION

Deep learning-based methods have enjoyed tremendous gains over the past years. While this has also been due to better architectures, a large part of this success is due to the ever-increasing size of datasets.

However, there are two major problems with the current datasets that the research community is increasingly becoming aware of and that could limit the progress in this domain if not addressed.

First, datasets are being revealed to be heavily biased. Works such as (Buolamwini & Gebru, 2018; Yang et al., 2020) find significant underrepresentation of women and those with darker skin in common datasets.

Biases in datasets, in turn, manifest as bias within models. For example, an image labelling algorithm from Google was found to label an image of two black people "Gorilla" (Simonite, 2018) and facial analysis models have been found to have significantly lower accuracy on darker female faces than on lighter male faces (Buolamwini & Gebru, 2018).

A second problem is the lack of consent for using these images to train AI models: As noted in (Birhane & Prabhu, 2021) while these datasets are collected under the Creative Commons licence, this license does not yield or say anything about their use in training AI models. This can potentially violate the licenses (as reproduction requires the ascription of the creator) since trained models can be probed to reveal whole training samples, such as addresses and bank accounts in GPT-2 (Carlini et al., 2020) or potentially even images (Orekondy et al., 2019). Therefore, there is a clear need for the removal of personally identifiable information within images. This would not only protect the individuals within the images (e.g. reduced risk of identify theft) but also ensure higher adherence to the General Data Protection Regulation (GDPR).

One common variable in both problems is the usage of raw, unedited data which tends to be either privacy infringing or biased. However, since collecting an unbiased dataset with consenting individuals would be very costly and potentially even unfeasible (Yang et al., 2020), in this paper we investigate approaches to mitigating these problems by modifying the current training data. Specifically, our contributions are as follows:

1. We investigate privacy-preserving measures on object detection performance.
2. We develop a novel method for balancing gender and race distribution in the training data while simultaneously removing personally-identifying information.
3. We test the fine-tuned models with two measures of model bias.

*Joint first authors. † :To whom correspondence should be addressed: yuki@robots.ox.ac.uk

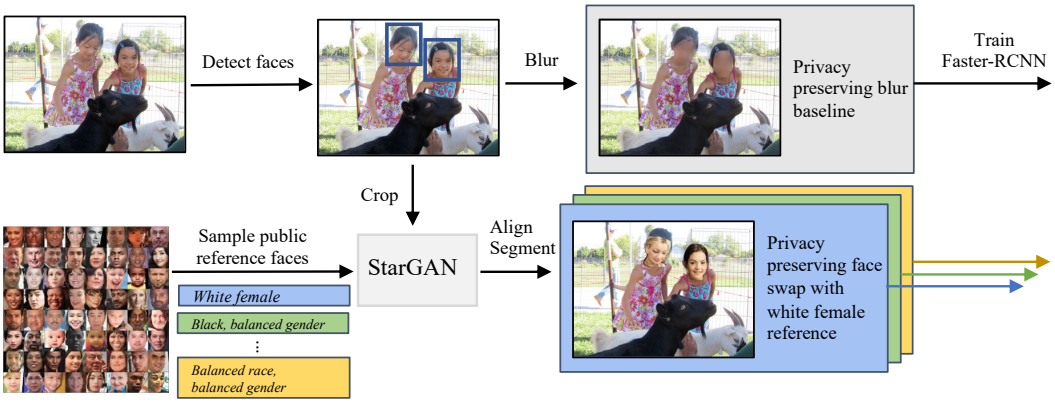

Figure 1: **Schematic overview of privacy-preserving approaches studied.** We explore various methods such as blurring faces and aligning and pasting faces fom a face dataset with known characteristics. Example image is taken from MS-COCO (ID=8690).

2 Related work

Measuring Bias. Experimentally, it has been shown that many models are biased: For example Buolamwini & Gebru (2018) show that face recognition tools perform worse for women of colour than white men, and in (Simonite, 2018; Kayser-Bril, 2020) publicly available models have been shown to be biased towards minorities.

Furthermore, Steed & Caliskan (2021) find that unsupervised models trained on ImageNet contain racial, gender, and intersectional biases. We use their test to probe bias in our models.

To mitigate biases, Zhang et al. (2017) develop a method to remove discriminatory effects of collected datasets prior to its statistical analysis.

Even with a balanced dataset, a model can amplify implicit gender biases, as shown by Wang et al. (2019), who propose to use a generative-adversarial approach to remove this information by painting over that part of the image. In this paper, we focus on how input data can be efficiently transformed to make datasets more balanced and potentially remove the association biases of stereotypes.

Preserving Privacy. Privacy is important within computer vision given the potential contradiction between using image recognition algorithms while at the same time limiting elements which expose identifiable and sensitive information. Preserving privacy in computer vision models has already been addressed to an extent using various methods, such as head inpainting (Sun et al., 2018), reducing resolution quality (Ryoo et al., 2017), and many more (Ren et al., 2018; Wu et al., 2018). Notably, (Orekondy et al., 2018) showed how to apply targeted obfuscation to areas of private information as to protect privacy yet also preserve the utility of the image. In (Yang et al., 2021), the authors find that blurring the faces in ILSVRC has a minimal impact on the accuracy of recognition models and the features learnt on face-blurred images transfer equally well to other tasks as those learnt on images with faces. However, so far little work has evaluated performances of preserving privacy on modern object detection algorithms.

Face Swapping. There has also been work in the area of transferring a face onto a target in a similar pose. Zhong et al. (2016) paste celebrity faces onto images to generate a synthetic dataset of novel (celebrity face, action) pair images.

Mitigating bias. In (Wang et al., 2019), the authors find that performance on multi-label prediction on COCO decreases only slightly when augmenting the model architecture with an adversarial loss to blur parts of the image that cause leakage of protected characteristics such as race and gender.

Figure 2: **Example Images.** Five versions of the same COCO image (ID=509864), each taken from one of our datasets. *Default* is the original, *Blurred* was created by blurring detected faces, and *White Male*, *Black & Balanced Gender*, and *Balanced Gender & Race* are created using the GAN-based method, augmentations outlined in Section 3.

3 METHODOLOGY

This paper explores how representations learnt for object detection change with various privacy-preserving augmentations of the training dataset. We augment the COCO dataset (Lin et al., 2014) by face-blurring and face-swapping using an adaptation of StarGAN (Choi et al., 2018). We use these, as well as the raw COCO dataset, to train Faster-RCNN (Ren et al., 2015). We measure the resulting model's object detection performance on both transformed and original versions. Finally, we attempt to measure the bias of the representations of all fine-tuned ResNet50 backbones via the Image Embedding Association Test (iEAT) (Steed & Caliskan, 2021).

Face Detection and Blurring. We augment the COCO dataset so that all faces are blurred and the face colours are scrambled. To blur the faces, we take an elliptical area inside each of the bounding boxes returned by the MTCNN face detector (Zhang et al., 2016) and perform a Gaussian blur (see Appendix for details). The intensity of the region of pixels inside each ellipse was shifted by a randomly sampled integer between -80 and +80. This provided a large range of colours, randomising the skin colours of faces in the image, while not saturating the intensity values.

GAN-based face swapping. We develop a pipeline for swapping faces in images using StarGAN (Choi et al., 2018), detailed in Figure 1. To make a swap, we take a crop of a face detection using MTCNN (Zhang et al., 2016) and sample a reference image from UTKFace (Song & Zhang, 2018) with a predefined distribution of gender and/or race[1]. To align the two faces, we first identify facial landmarks with FAN (Bulat & Tzimiropoulos, 2017), then generate a matrix transformation to map landmarks to fixed coordinates. Using StarGAN, we create an artificial face that has the style of the source face (facial expression) and the texture of the reference face (gender, skin and hair colour). Finally, we warp the generated face to align its landmarks with the face in the source image, remove the background using a pretrained instance segmentation model (Mask-RCNN (He et al., 2017)), and paste the segmented face in the original image. The detailed procedure is in A.3.

Bias measure: iEAT test. To measure the bias of image representations, we use the iEAT test (Steed & Caliskan, 2021), which is adapted from the social psychology IAT test (Greenwald et al., 1998). The test measures the differential association of some target concepts (e.g. "woman" vs "man") with a set of attributes (e.g. "maths" vs "arts"), in all images (a total of 697). We first extract the representations of these visual stimuli using the ResNet backbone in Faster-RCNN models trained on the different datasets. Then, we calculate the cosine distance between the normalized representations of different visual stimuli, and record the p-value and effective size d of the null hypothesis significance testing whether the model is biased.

Privacy-Attribute Leakage: Linear classifier. We say a model "leaks" gender data if it learns different representations for "man" and "woman" when there has only been label "person" in the training set. To measure leakage, we train a linear binary classifier on top of the representations of frozen ResNet50 backbones from the Faster-RCNNs. We test leakage using two datasets, framewise PA-HMDB51 (Wu et al., 2020) and a Kaggle gender classification dataset (Orsolini, 2019). We use this as a proxy for a model's performance on out-of-distribution datasets, and in particular its ability to pick up previously unseen attributes (e.g. the "blurred" model has not seen faces).

[1]Definition from UTKFace as in the US census, classified as White, Black, Asian, Indian, and Others

4 RESULTS

Table 1: Object detection performance on COCO and the augmentations of COCO we make. We present mean AP, AP_{75} and AP♀ (for mAP of person category) of Faster-RCNN FPN trained (1x schedule) and evaluated on the different combinations datasets.

	Eval	**Original**			**Blurred**			**Black F.**			**Bal. G.& R.**		
	Training	AP	AP_{75}	AP ♀	AP	AP_{75}	AP ♀	AP	AP_{75}	AP ♀	AP	AP_{75}	AP ♀
(a)	Original	37.9	41.0	52.5	37.5	40.5	51.5	37.6	40.7	52.1	37.9	41.0	51.5
(b)	No Faces	36.5	39.4	49.4	36.2	39.2	48.4	36.5	39.4	49.4	36.5	39.4	49.4
(c)	No Persons	29.7	32.8	0.0	29.6	32.7	0.0	29.7	32.7	0.0	29.7	32.7	0.0
(d)	Blurred	37.8	41.1	52.5	37.6	**40.9**	**52.4**	37.8	41.0	52.5	37.8	41.1	52.6
(e)	White, M	37.9	**41.2**	52.6	37.7	**40.9**	52.1	**37.9**	**41.2**	52.6	37.9	**41.2**	52.6
(f)	Black, F	37.9	41.0	52.6	**37.7**	40.6	51.5	37.8	40.8	52.5	37.9	41.0	52.6
(g)	Black, Bal. G.	37.9	41.0	52.7	37.5	40.4	51.8	**37.9**	41.0	52.8	37.9	41.0	**52.8**
(h)	Bal. G & R	**38.0**	**41.2**	**52.8**	37.6	40.8	51.7	**37.9**	**41.2**	**52.8**	**38.0**	**41.2**	**52.8**

Object Detection Performance The original COCO 2017 dataset contains 118,287 training images. Removing all images with detected faces yields 86,721 training images, while removing all images labelled with the "person" class leaves a training set of only 54,172 images. These drops in the number of available training are also echoed in much lower performances in Tab. 1(b-c). Compared to this, even our simple baseline, blurring of faces (row d), works much better and also does not suffer from a loss in performance when evaluating on blurred faces. Overall, all models trained on modified datasets perform as well as the default model when detecting objects in the COCO dataset. There is no significant drop in AP (<0.1) across any model and measure. Each model trained on an modified dataset performs better on that dataset than the default model. In particular, we find the model with balanced gender and race (row h) to perform well across the board. This can be attributed to the artifacts around modified faces that the models trained on the modified datasets pick up. However, models trained on any modified dataset perform no worse on the default dataset, where no face obfuscation has been performed, than the default model. We conclude that given a suitable face obfuscation method, we can train an object detector that does not degrade in performance on real data, despite never having seen a real face.

iEAT Bias Measurements The results from the iEAT association test show only minor differences between the default COCO model and our augmentations. The model *Black Female* shows a 20% decrease in bias against "Arab-Muslim" and *Balanced Gender & Race* shows a minor increase in bias against "Disability". Biases concerning "(modern) weapons" are statistically significant in *Balanced Gender* and *Balanced Gender & Race* even though they are not in other models. All other differences are either (i) of negligible magnitude (e.g. "Weight"), (ii) are differences of effect size in statistically insignificant results (e.g. in sexuality) or (iii) are differences in the empirically untested intersectional categories (see A.5). This, in combination with the highly significant biases of a randomly initialised RCNN, cast doubt on the validity of the iEAT test.

Table 2: Selected iEAT results for effective size. Original is the RCNN model trained on original COCO; COCO Blurred is the RCNN model trained on COCO with faces blurred; Bal. Gender&Race is the RCNN model trained on COCO with modified faces balanced along gender and racial dimensions; We also report the ResNet trained on ILSVRC-12 with faces blurred (Yang et al., 2021) and a randomly initialized network.

Model trained on	Weight	Disability	Arab-Muslim	Gender-Science
Random	1.8275^{***}	-0.5674	0.4497	0.0690
ILSVRC-12 Blurred Faces	1.2277^{***}	1.3358^{**}	0.8474^{**}	0.4691^{**}
Original COCO	1.4469^{***}	1.0590^{*}	0.7595^{**}	0.3471^{*}
COCO Blurred	1.4563^{***}	1.0402^{*}	0.7484^{**}	0.3758^{**}
Bal. Gender&Race	1.8793^{***}	-0.8023^{*}	0.5498^{**}	-0.2099^{*}

*: p<0.1; **: p<0.05; ***: p<0.01

Privacy-Attribute Leakage Results The accuracy of gender classifiers trained on all models was similar. This suggests that there is no significant reduction in the possibility of decoding gender from the features, even in our backbone *Blurred* which never saw a face during final training. It is likely that cues from a person's hair or body must also be altered to successfully obfuscate gender. While obfuscating gender is not a necessary condition for removing bias, we believe further work is required. Another next step involves pretraining only the weights of ResNet backbones with manipulated input data before finetuning on COCO object detection and repeating these experiments.

Table 3: Selected leakage results.

Model	PA-HMDB51	M/F
Original	80.6 ± 1.0	85.4 ± 0.8
Blurred	80.3 ± 0.6	85.7 ± 1.5
Bal. G.&R.	80.7 ± 0.7	85.4 ± 0.7

5 Conclusion

Does a dataset need real human faces? Not for object detection. In this paper, we show that at least for finetuning on COCO, this is not the case. We find that even when faces are blurred or edited to be more balanced in terms of represented gender and race, the object detector learns just the same. We also investigate whether it is possible to de-bias the models using two measures proposed. However, here we report negative and inconsistent findings that at least partially point to potential shortcomings of these measures tested.

Acknowledgements

The OxAI society is grateful for support from Google Academic Research Credits program.

References

Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In International Conference on Computer Vision, 2017.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research, pp. 77–91, 2018.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. 2020.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology, 74(6):1464, 1998.

K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988, 2017. doi: 10.1109/ICCV.2017.322.

Mike Hintze. Viewing the GDPR through a de-identification lens: a tool for compliance, clarification, and consistency. International Data Privacy Law, 2017. ISSN 2044-3994. doi: 10.1093/idpl/ipx020.

Nicolas Kayser-Bril. Google apologizes after its vision ai produced racist results, 2020. URL `https://algorithmwatch.org/en/story/google-vision-racism/`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2017.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), ECCV, 2014. ISBN 978-3-319-10602-1.

Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

Tribhuvanesh Orekondy, Seong Joon Oh, Yang Zhang, Bernt Schiele, and Mario Fritz. Gradient-leaks: Understanding and controlling deanonymization in federated learning. In NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality, 2019.

Johnathan Orsolini. Men/women classification: A jpg dataset for male/female classification, 2019. URL https://www.kaggle.com/playlist/men-women-classification.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, 2015.

Zhongzheng Ren, Yong Jae Lee, and Michael S. Ryoo. Learning to anonymize faces for privacy preserving action detection. In Proceedings of the European Conference on Computer Vision (ECCV), 2018.

Michael S. Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 4255–4262, 2017.

Tom Simonite. When it comes to gorillas, google photos remains blind, 2018. URL https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.

Yang Song and Zhifei Zhang. Utkface: Large scale face dataset, 2018. URL https://susanqq.github.io/UTKFace/.

Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. CoRR, abs/2010.15052, 2021.

Qianru Sun, Liqian Ma, Seong Joon Oh, L. Gool, B. Schiele, and M. Fritz. Natural and effective obfuscation by head inpainting. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5050–5059, 2018.

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. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

Z. Wu, H. Wang, Z. Wang, H. Jin, and Z. Wang. Privacy-preserving deep action recognition: An adversarial learning framework and A new dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 606–624, 2018.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 547–558, 2020.

Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet, 2021.

K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 2016.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI'17, 2017.

Yujie Zhong, Relja Arandjelovic, and Andrew Zisserman. Faces in places: compound query retrieval. In Procedings of the British Machine Vision Conference 2016, 2016.

A APPENDIX

A.1 DISCUSSION ON GDPR

Under the General Data Protection Regulation (GDPR) biometric data is considered a special category of data (Art. 9.1) and facial images are classified as a type of 'biometric data' (Art 4.14). As such it would be reasonable to conclude that the storage of facial images would likely fall under the GDPR. However, as we have shown, this paper's method is able to remove certain personally identifiable information (PII), leading to what is known as a 'de-identification' process (Hintze, 2017). By removing PII, whether the data falls under the GDPR and, for example the storage limitation principle (Art. 5.1.e), is brought into question. We believe that as our method reduces PII from images, it also reduces the regulatory requirements for those storing images. This is dependent on if an argument can be made whether the method anonymizes or pseudonymizes the images. If it anonymizes the images, then they would no longer fall under the GDPR (see Recital 26). However, if it pseudonymizes them then it would fall under the GDPR (Art. 4.3b). Regardless of the extent to which our method 'de-identifies' subjects, we believe that by reducing PII this provides greater flexibility for researchers to work with computer vision models while adhering to 'data minimisation' and 'storage limitations' principles where applicable. Therefore, as our method retains information such as facial expression and (to some extent) photorealism it is more favorable than ablation or blurring. We believe that large available datasets could be processed using similar methods before public release to minimise privacy and bias concerns. However, given that under the law it often depends on the specific case in order to provide a legal argument, the fact remains that our method of protecting privacy and preserving performance seems to be of beneficial use.

A.2 BLURRING DETAILS

We apply Gaussian blurring with kernel length $h/6$ (where h is the height of the original bounding box) and a standard deviation of 20 pixels.

A.3 FACE SWAPPING WITH STARGAN

Suppose we want to swap a face in image I. We crop out the original face F which can be decomposed as $F = (A, E, L, B)$, where A is appearance and texture, E is the style and facial expressions, L is position of landmarks, and B is background behind the face. Given a predefined gender and/or race setting p for the target dataset, we choose a reference face $R_p = (A'_p, E', L', B')$. To align the two faces, we calculate two alignment matrices $\mathcal{M}$, $\mathcal{M}'$ (calculated such that $\mathcal{M}L = L_0$ and $\mathcal{M}'L' = L_0$, where L_0 is a fixed face framework). We apply these to get two aligned face that we denote $F^{(a)}$ and $R^{(a)}$ by $F^{(a)} = \mathcal{M}F,\ R_p^{(a)} = \mathcal{M}'R_p$. Using StarGAN ($\mathcal{T}$), we merge the two faces to create $G_p^{(a)} = \mathcal{T}(F^{(a)}, R_p^{(a)}) = (A_p'^{(a)}, E^{(a)}, L_0, B'^{(a)})$, which retains the style of the original face $E^{(a)}$ but the appearance and background of the reference face $A_p'^{(a)}B'^{(a)}$. The synthesised face is then reverted back to the original position by using an inverted alignment matrix ($G_p = \mathcal{M}^{-1}G_p^{(a)} = (A'_p, E, L, B')$). The background of the artificial face is removed by segmentation (denoted by $\mathcal{S}$) to create $S_p = \mathcal{S}G_p = (A'_p, E, L, \emptyset)$. Finally, the face is pasted back to the original image, giving us I_p, an altered image with desired gender and skin colour.

Hyper-parameters of StarGAN: Most hyper-parameters are default values from the pretrained models: MTCNN, FAN, StarGAN, and Mask-RCNN.
We only change the margin parameter of StarGAN to 0.8 and the threshold of MTCNN to 0.5.

A.3.1 GENDER CLASSIFICATION

Male/Female Kaggle dataset We train the frozen ResNet50 backbones for 20 epochs, using Adam optimizer (Kingma & Ba, 2017) with learning rate 0.001 and a batch size of 16, and use 3-fold cross validation.

PA-HMDB51 dataset PA-HMDB51 contains videos labelled by actions (such as "brush_hair") and frame-by-frame labels of privacy attributes including skin colour and gender. We extract frames containing only one "male"/"female" gender label to create a dataset of images labelled by gender

and use this to train a gender classifier. As with the Male/Female Kaggle dataset, we train the frozen ResNet50 backbones for 1 epoch, using Adam optimizer (Kingma & Ba, 2017) with learning rate 0.001 and a batch size of 16, and use 3-fold cross-validation.

A.4 LEAKAGE RESULTS

Table A.1: Full leakage results.

Model	PA-HMDB51	M/F
Original	80.6 ± 1.0	85.4 ± 0.8
Blurred	80.3 ± 0.6	85.7 ± 1.5
White M	80.2 ± 0.7	85.7 ± 0.6
Black F	79.9 ± 0.6	85.8 ± 0.3
Bal. G.&R.	80.7 ± 0.7	85.4 ± 0.7

A.5 IEAT TEST RESULTS

The iEAT test measures bias along a number of variables, but we focus on the ones that we identify as *social* biases, and only discuss statistically significant results. There is a decrease in the *Gender-Science*, *Arab-Muslim* and *Weight* categories when comparing **Balanced-intersectional** to **Default**.Comparing **Black-female** to **Default**, we see a 10% to 20% decrease of effective size in *Gender-Science*, *Arab-Muslim* and *Disability*. **White-male** increases the bias in all categories with significant results, *Weight* and *Disability*, compared to **Default**. Similarly, **Blurred** increases *Gender-Science* and *Weight* biases, while slightly decreasing along the *Disability* and *Arab-Muslim* categories.

For completeness, we include results for a randomly initialised RCNN, as well as an RCNN backbone pretrained on ImageNet. While the results of the randomly initialised one can not be interpreted, the ImageNet-pretrained one shows the smallest bias in the *Weight* and *Arab-Muslim* categories, and the highest in *Disability*. Models trained on altered datasets show similar bias results, suggesting that data preprocessing has a small effect. Changing the layer from which we take representations extraction layer shows no difference in the results

An important thing to note is that as mentioned in the original iEAT paper (Steed & Caliskan, 2021), the IAT bias test on humans, where this test is adapted from, did not include the bias tests on intersectionality. Therefore, the results for intersectional bias have not been empirically compared and supported. The fact that even the randomly initialised model has significant biases in some categories means that the intersectional biases are likely false positives.

In addition, the tests use a small number of visual stimuli (normally less than 10 for each attribute and concept) and sample 3000-10000 partitions from the union of 2 attributes and 2 concepts, so statistical bias could result from the small selection of images.

Table A.2: iEAT bias test full results: p value and effective size D. Abbreviations: **Rnd**: Randomly initialised, **IDf**: Imagenet pretrained, **Df**: default COCO pretrained, **Bl**: Blurred, **BF**: Black female, **WM**: White male, **Bal-R**: balanced along race, **Bal-G**: balanced along gender, **Bal-I**: balanced intersectionally.

p-value / Effect-size

	Rnd	IDf	Df	Bl	BF	WM	Bal-R	Bal-G	Bal-I
Insect-Flower	1.000 -0.945	0.988 -0.553	1.000 -0.970	1.000 -1.125	1.000 -0.963	1.000 -1.111	1.000 -1.086	1.000 -1.033	1.000 -1.030
Gender-Science	0.378 0.069	0.219 0.177	0.060* 0.347	0.046** 0.376	0.084* 0.313	0.119 0.267	0.133 0.250	0.051* 0.369	0.065* 0.337
Gender-Career	0.133 0.253	0.081* 0.305	0.735 -0.140	0.779 -0.174	0.746 -0.154	0.848 -0.234	0.781 -0.172	0.666 -0.102	0.798 -0.196
Disability	0.786 -0.567	0.014** 1.285	0.071* 1.059	0.071* 1.040	0.071* 0.974	0.071* 1.107	0.071* 1.101	0.071* 1.056	0.071* 1.129
Asian	0.118 0.740	0.233 0.483	0.325 0.294	0.310 0.335	0.022** 1.147	0.529 -0.061	0.172 0.593	0.563 -0.093	0.452 0.080
Arab-Muslim	0.158 0.450	0.089* 0.620	0.048** 0.760	0.044** 0.748	0.084* 0.628	0.100 0.595	0.108 0.567	0.050** 0.746	0.049** 0.754
Age	0.963 -1.000	0.343 0.277	0.513 -0.022	0.543 -0.096	0.530 -0.057	0.538 -0.064	0.472 0.033	0.502 -0.003	0.447 0.087
Weight	<1e-3*** 1.828	0.003*** 1.117	<1e-3*** 1.447	<1e-3*** 1.456	<1e-3*** 1.493	<1e-3*** 1.523	<1e-3*** 1.358	<1e-3*** 1.484	<1e-3*** 1.410
Weapon (Modern)	0.168 0.568	0.181 0.579	0.384 0.180	0.723 -0.374	0.127 0.672	0.795 -0.515	0.489 0.035	0.037** 1.026	0.035** 1.010
Weapon	0.421 0.118	0.981 -1.138	0.120 0.700	0.215 0.477	0.374 0.192	0.740 -0.418	0.427 0.110	0.457 0.062	0.707 -0.347
Skin-Tone	<1e-3*** 1.531	0.446 0.075	0.355 0.151	0.341 0.108	0.395 0.087	0.302 0.163	0.339 0.198	0.385 0.076	0.377 0.101
Sexuality	0.371 0.163	0.694 -0.244	0.606 -0.137	0.558 -0.084	0.565 -0.075	0.583 -0.100	0.532 -0.039	0.628 -0.147	0.644 -0.178
Religion	0.217 0.437	0.338 0.229	0.206 0.455	0.276 0.337	0.269 0.344	0.222 0.429	-0.292 0.322	0.227 0.412	0.259 0.358
Race	0.021** 1.137	0.933 -0.856	0.648 -0.268	0.635 -0.306	0.642 -0.252	0.669 -0.324	0.705 -0.384	0.680 -0.352	0.649 -0.271
Native	0.489 0.021	0.998 -1.323	0.821 -0.474	0.634 -0.174	0.862 -0.568	0.793 -0.419	0.706 -0.270	0.885 -0.613	0.430 0.094
Intersectional-Valence-WMWF	0.032** 0.585	0.045** 0.533	0.017** 0.661	0.011** 0.708	0.040** 0.554	0.038** 0.561	0.041** 0.557	0.014** 0.683	0.012** 0.711
Intersectional-Valence-WMBM	<1e-3*** 1.052	0.581 -0.072	0.257 0.215	0.375 0.103	0.380 0.095	0.280 0.188	0.315 0.163	0.312 0.163	0.134 0.354
Intersectional-Valence-WMBF	<1e-3*** 1.163	0.313 0.156	0.139 0.352	0.237 0.234	0.156 0.323	0.133 0.354	0.172 0.303	0.191 0.288	0.109 0.391
Intersectional-Valence-WFBM	<1e-3*** 1.313	0.044** 0.536	0.007*** 0.736	0.019** 0.648	0.038** 0.555	0.023** 0.621	0.026** 0.604	0.012** 0.704	0.002*** 0.927
Intersectional-Valence-WFBF	<1e-3*** 1.361	0.007*** 0.765	<1e-3*** 0.987	0.002*** 0.878	0.004*** 0.836	0.003*** 0.828	0.005*** 0.817	<1e-3*** 0.927	<1e-3*** 1.080
Intersectional-Valence-FM	0.459 0.022	0.256 0.150	0.134 0.252	0.128 0.252	0.263 0.146	0.220 0.173	0.208 0.187	0.129 0.259	0.071* 0.327
Intersectional-Valence-BW	<1e-3*** 1.224	0.080* 0.319	0.006*** 0.558	0.021** 0.457	0.024** 0.442	0.013** 0.494	0.018** 0.462	0.011** 0.511	0.001*** 0.678
Intersectional-Valence-BFBM	0.845 -0.338	0.787 -0.257	0.582 -0.068	0.603 -0.086	0.712 -0.181	0.628 -0.113	0.606 -0.091	0.583 -0.072	0.460 0.030
Intersectional-Gender-Science-WMWF	0.980 -0.649	0.957 -0.539	0.850 -0.333	0.548 -0.041	0.781 -0.243	0.817 -0.295	0.828 -0.298	0.874 -0.372	0.889 -0.396
Intersectional-Gender-Science-WMBM	0.006*** 0.768	0.424 0.058	0.135 0.353	0.474 0.018	0.626 -0.109	0.521 -0.018	0.320 0.155	0.221 0.249	0.116 0.377
Intersectional-Gender-Science-WMBF	<1e-3*** 1.017	0.125 0.365	0.042** 0.548	0.225 0.253	0.141 0.357	0.101 0.418	0.042** 0.543	0.098* 0.410	0.032** 0.577
Intersectional-Gender-Science-MF	0.742 -0.155	0.685 -0.108	0.594 -0.057	0.316 0.108	0.327 0.104	0.354 0.081	0.436 0.038	0.649 -0.084	0.603 -0.066
Intersectional-Gender-Career-WMWF	0.950 -0.524	0.688 -0.162	0.926 -0.463	0.862 -0.352	0.833 -0.315	0.871 -0.366	0.874 -0.368	0.933 -0.474	0.939 -0.487
Intersectional-Gender-Career-WMBM	0.877 -0.372	0.275 0.194	0.076* 0.456	0.045** 0.539	0.122 0.375	0.045** 0.528	0.119 0.383	0.207 0.269	0.145 0.342
Intersectional-Gender-Career-WMBF	0.019** 0.659	0.874 -0.365	0.667 -0.141	0.704 -0.172	0.526 -0.021	0.735 -0.204	0.628 0.628	0.498 0.008	0.521 -0.009
Intersectional-Gender-Career-MF	0.734 -0.147	0.743 -0.145	0.618 -0.069	0.476 0.010	0.463 0.020	0.521 -0.008	0.577 -0.043	0.675 -0.102	0.638 -0.080

A.6 EXAMPLE-IMAGE GALLERY

Figure A.1: A selection of example images used for training. They are labelled by their COCO IDs, such as "32901", as well as the dataset they belong to. Details about the creation of each dataset can be found in 3.