Attention IoU: Examining Biases in CelebA using Attention Maps

Aaron Serianni Tyler Zhu Olga Russakovsky Vikram V. Ramaswamy
Princeton University
{serianni, tylerzhu, olgarus, vr23}@princeton.edu

Abstract

Computer vision models have been shown to exhibit and amplify biases across a wide array of datasets and tasks. Existing methods for quantifying bias in classification models primarily focus on dataset distribution and model performance on subgroups, overlooking the internal workings of a model. We introduce the Attention-IoU (Attention Intersection over Union) metric and related scores, which use attention maps to reveal biases within a model's internal representations and identify image features potentially causing the biases. We analyze the CelebA dataset, finding that Attention-IoU uncovers correlations beyond accuracy disparities. Through an investigation of individual attributes through the protected attribute of `Male`*, we examine the distinct ways biases are represented in CelebA. Lastly, by subsampling the training set to change attribute correlations, we demonstrate that Attention-IoU reveals potential confounding variables not present in dataset labels. Our code is available at* https://github.com/aaronserianni/attention-iou.

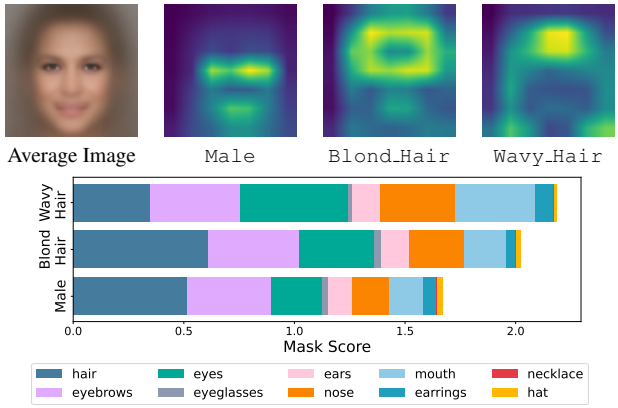


Figure 1. We use attention maps to understand which image regions a model relies on for the target classification task. Our proposed Attention-IoU framework provides insights into how models represents biases between correlated attributes. For example, consider the spatially related attributes of blond and wavy hair in the CelebA dataset [15], which have similarly correlations to the `Male` label. They are attended to differently by the model, with blond hair being more related to `Male` in both average attention map (*top*) and the Attention-IoU mask score (*bottom*).

1. Introduction

Biases in computer vision models can lead to failures in model performance and unequal behavior for different groups. These biases are often caused by spurious correlations, where a model relies on an attribute that is associated with, but not causally related to, the target. A model dependent on such spurious correlations might then perform poorly on out-of-distribution test data or exhibit low accuracy for groups for which the correlation does not hold. This becomes more concerning for tasks involving people, since these correlations can cause models to discriminate against societally protected groups such as gender, race, age, ethnicity, and income [3, 5, 9, 25, 32, 33].

Past works have extensively investigated biases and spurious correlations through the lens of dataset labeling and model accuracy. For example, fairness metrics reveal disparities in model accuracy between groups or individuals [4, 16, 18, 27]. Others have created tools to surface biases by analyzing and categorizing objects, gender, skin tone, geographical labels, among others, sometimes in combination with model predictions [2, 29].

However, these approaches are often limited by the labels present within the dataset, only able to find biases at a coarse level. For example, while these metrics excel at identifying when the classification of a person's attributes might depend on gender, they are unable to highlight the specific features of the person's gender presentation that the model uses to make a prediction. In the absence of fine-grained labels, interpretability methods like attention maps [23, 31, 34] hold the potential to reveal representations of correlations within a model, and how they might affect the model's output.

In this paper, we propose *Attention-IoU*, a generalized intersection-over-union metric that uses attention maps to measure biases in image classification models. We specifically aim to quantify spurious correlations for when a model

relies on regions of images that are not directly relevant to the target classification tasks. For example, within the CelebA dataset [13, 15] of people's faces, blond hair is correlated with a person being labeled not male. As such, a model trained to identify the 'blond hair' attribute may use gendered aspects of people's faces in addition to using hair features. Thus, the model may attend to regions such as the eyes, nose, and mouth of people as well as their hair (Fig. 1).

We examine CelebA because the dataset is a widely-used evaluation benchmark for fairness methods, spanning dataset bias identification to model debiasing [11, 17, 19, 20, 22, 24, 30]. With CelebA, we demonstrate that Attention-IoU can identify specific ways in which the protected `Male` attribute might influence other attributes. We also show that attributes can be unevenly influenced by the classifier's representation of the protected `Male` attribute, and that certain attributes have biases beyond simple correlations in dataset labels. These insights reveal ways in which computer vision models might be biased, allowing the community to develop better debiasing techniques.

2. Method

Existing bias metrics for classification models focus on how the models perform with respect to certain groups within a dataset [7, 27, 33]. These common approaches often only consider the final predictions of models, but in line with other works [1, 6, 12, 14], we aim to understand *why* these biases might occur. The key insight for our bias identification method is the following: if a model learns a spurious correlation between a target attribute and a confounding attribute in the dataset, it will learn to use features helpful for the *confounding attribute* instead of the target attribute. This lets us quantify bias by comparing a model's attention map for the target attribute to either attention maps of confounding attributes or ground-truth feature maps.

Attention Map Metrics. We use Gradient-weighted Class Activation Mapping (GradCAM) to obtain attention maps for target attributes [23]. Given an input image $\mathbf{x}$ and target attribute a, let the attribute-specific attention map be $\mathrm{GradCAM}_a(\mathbf{x})$. The metric should be able to compare two real-valued attention maps with each other, as well as an attention map with a binary ground-truth feature mask. Based on these constraints, we propose a generalized IoU metric, which we refer to as *Attention-IoU*, that works on weighted dense-pixel maps and is size and scale invariant. Given two maps $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{h \times w}$, which can be either attention maps or feature masks, denote their L_1 normalized maps as $\widehat{\mathbf{M}}_i = \mathbf{M}_i / \|\mathbf{M}_i\|_1$, which are akin to probability density functions. The metric is defined as

$$\mathcal{B}_{\text{A-IoU}}(\mathbf{M}_1, \mathbf{M}_2) = \frac{\langle \widehat{\mathbf{M}}_1, \widehat{\mathbf{M}}_2 \rangle_F}{\left\| \frac{\widehat{\mathbf{M}}_1 + \widehat{\mathbf{M}}_2}{2} \right\|_F^2} = \frac{\sum_{i,j} (\widehat{\mathbf{M}}_1)_{ij} \cdot (\widehat{\mathbf{M}}_2)_{ij}}{\sum_{ij} \left(\frac{\widehat{\mathbf{M}}_1 + \widehat{\mathbf{M}}_2}{2} \right)_{ij}^2}$$

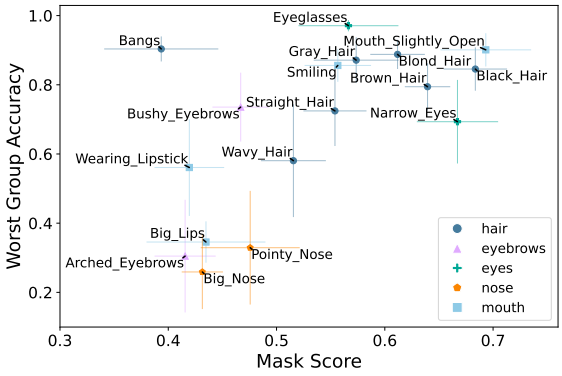


Figure 2. **Evaluation of mask score using GradCAM on CelebA test set with attribute-specific feature masks, compared to worst group accuracy with `Male`**. Groups are considered based on ground-truth labels for the different combinations of target attribute and `Male`. If the number of images in a group is less than 1% of the test set, the group was excluded from consideration.

where $\langle \mathbf{A}, \mathbf{B} \rangle_F$ is the Frobenius inner product, *i.e.*, the sum of the element-wise matrix product, and $\|\mathbf{A}\|_F^2$ is the Frobenius norm, *i.e.*, the sum of squared entries of the matrix.

Bias Scores. Using Attention-IoU, we define two methods to score biases in a model for a given target. The *heatmap score* compares the attention map for the target attribute with the attention map of a chosen protected attribute using $\mathcal{B}_{\text{A-IoU}}$. The *mask score* is computed between the target's attention map and a chosen ground-truth feature mask corresponding to the input image. As the size of the attention map is the size of the final convolution layer, whereas the feature mask is the size of the input image, the feature mask is downsampled with bilinear interpolation. The heatmap and mask scores are averaged over all images in a given set.

Advantages of Attention-IoU. Attention-IoU has several advantages over existing bias detection methods. First, since the metric is based on attention maps, it highlights specific regions of the sensitive attribute that most contribute to the target attribute prediction. Thus, we are able to identify bias at a more fine-grained level than other bias metrics. Next, by visualizing the scores separately for different types of images, we can infer if the bias is different for the different sets. For example, this allows us to understand if the features of the sensitive attribute are used solely when the attribute takes on a particular value. Finally, the metric allows us to unearth potential confounding variables; *i.e.*, when the bias is due to more than the simple proportion of labels within the training dataset.

3. Analyzing CelebA

In this section, we analyze the CelebA dataset [15] using Attention-IoU. CelebA is labeled with 40 different

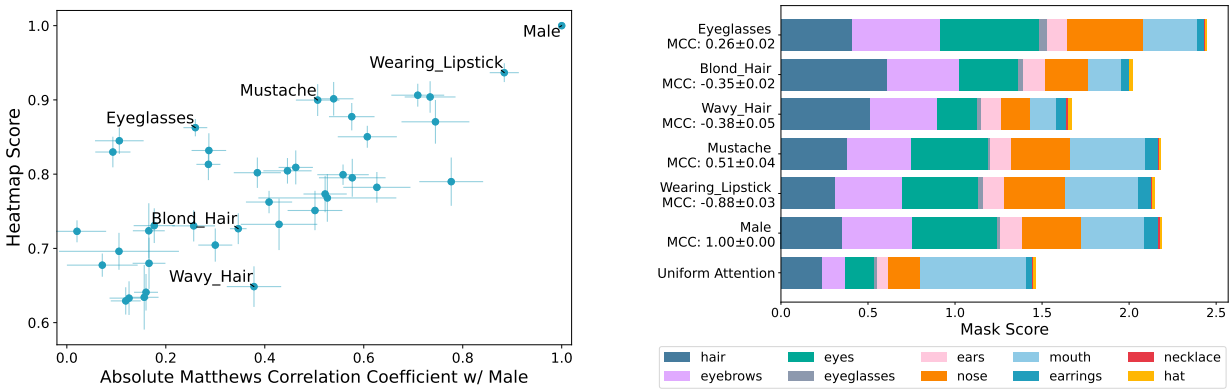


Figure 3. **Comparison of attributes with the `Male` attribute heatmap**. (*Left*) We compare Attention-IoU with the absolute value of the Matthews correlation coefficient between the predictions of the attribute and `Male`, noticing a strong positive trend. Some attributes are outliers to this trend, including `Eyeglasses` and `Mustache`, which lie above this trend, and `Wavy_Hair`, which lies below. (*Right*) We measure the mask score for a selection of attributes. We notice that the heatmap for `Male` attends most strongly to the eye, eyebrows, and mouth region, which is closely mimicked by `Wearing_Lipstick`. We can also compare attributes like `Blond_Hair` and `Wavy_Hair`, and find that the main difference between their heatmaps is in the eye region.

attributes including both attributes localized to specific face regions (*e.g.*, `Big_Nose`, `Mouth_Slightly_Open`, `Blond_Hair`) and attributes that are more global (*e.g.*, `Male`[1], `Heavy_Makeup`). We use Attention-IoU to understand more about the attributes present in the dataset, and how they might influence each other.

We start by evaluating heatmaps using ground-truth masks, for attributes that are localized and have associated masks. We choose a subset of 17 CelebA attributes that have directly corresponding feature masks, and calculate the respective mask score for each attribute (Fig. 2). There is not a strong correlation between worst group accuracy (WGA) and the mask score. This is not surprising, since dataset bias is not immediately correlated to a singular attribute's labeling. Instead, an attribute's WGA and bias is dependent on the features in the image and the distribution of its label with the labels of other attributes.

Comparison with the `Male` heatmap. In line with prior works, which investigate the impact of bias due to the protected `Male` attribute, we next examine the correlation between the heatmaps of different attributes and the heatmap for the `Male` attribute. We compute Attention-IoU for all 40 attributes with `Male` (Fig. 3 *left*). We measure the correlation between the attribute and the `Male` label using the absolute value of Matthews correlation coefficient (MCC), which is tailored for comparing two binary variables. There is a clear positive trend between the heatmap score and predicted label MCC. Some attributes are outliers to this trend, such as `Mustache` and `Eyeglasses` having higher heatmap scores, and `Wavy_Hair` having a lower heatmap score. We also report the mask score for selected attributes (Fig. 3 *right*). The mask score for `Male` demonstrates that the models attend most strongly to the eye, eyebrow, and mouth region of the face, and slightly less to the nose and hair regions. We notice that this is most closely replicated by `Wearing_Lipstick`, validating the high heatmap score. This per-region score computation also allows us to understand *how* features of different attributes differ: for example, the main difference between `Blond_Hair` and `Wavy_Hair` appears to be in how much the models attend to regions around the eyes and nose. We now analyze in detail four attributes representative of those with distinct properties.

Wearing Lipstick. `Wearing_Lipstick` has the highest absolute correlation with `Male` out of all 40 attributes, with an MCC of 0.88 ± 0.03. Furthermore, this correlation is predictive in both directions. One would expect that the attention map for `Wearing_Lipstick` would highlight the mouth region. However, the mask score shows that the models attend to the eyes, eyebrows, nose, and hair regions, in addition to the mouth. In fact, the mask score distribution for `Wearing_Lipstick` is closely similar to that of `Male`, only with a slightly higher mouth mask score. This close similarity between `Wearing_Lipstick` and `Male` is reflected in the heatmap score, the highest of any attribute.

Eyeglasses. `Eyeglasses` is an outlier to the heatmap score trend, having significantly higher heatmap scores compared to other attributes with similar MCCs. The at-

[1]We acknowledge that these binary feature labels in CelebA, especially the `Male` label, forces people's presentations to fit into binaries. The `Male` label inherently assumes that an individual's gender presentation is tied to their gender identity. It is not clear what standards the creators of CelebA use in their definition of the `Male` label and other feature labels. However, for our goal of creating and evaluating bias metrics, we follow existing literature in our use of CelebA labels.

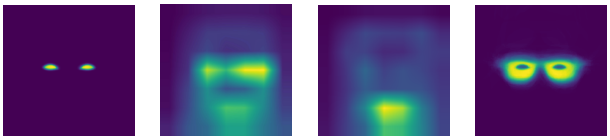


Figure 4. **Average heatmaps for `Male`**. We train models to predict `Male` when `Eyeglasses` are absent (*center-left*) and present (*center-right*). We notice a stark difference in the heatmaps suggesting that the features used by the model for predicting `Eyeglasses` is different from those being used to predict `Male`, despite them being co-localized in the original models.

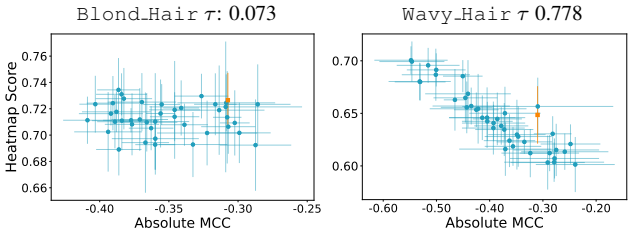


Figure 5. **Varying the correlation in the training dataset**. To understand if the correlations are indeed responsible for the mask scores, we subsample the dataset to vary the ground-truth MCC between `Blond_Hair` and `Wavy_Hair` and `Male`. We find that changing the ground-truth MCC for `Blond_Hair` (*left*) does not change the heatmap score, while changing the MCC for `Wavy_Hair` (*right*) results in a strong change in the heatmap score (orange/square indicates the original results). This suggests that there might be a hidden confounder present between `Blond_Hair` and `Male`, leading to the large heatmap score.

tribute is moderately correlated with `Male`, having an MCC of 0.26 ± 0.02, suggesting that `Male` is unlikely to influence the prediction of `Eyeglasses` much (or vice versa). As shown by the `Eyeglasses` mask score, the models attend strongly to the eyes, eyebrows, and nose regions. Surprisingly for an attribute with a low MCC, the heatmap score for `Eyeglasses` is high at 0.86 ± 0.01. We posit that this might be due to one of the weaknesses within Attention-IoU: it's unable to detect when features are co-localized. In this case, we notice in Fig. 3 (*right*) the heatmap attends highly to eyes and eyebrows, similar to that in `Male`.

To verify, we train two models to classify the `Male` attribute, one with just images for which `Eyeglasses` are present, and another for which `Eyeglasses` are absent. We hypothesize that if the `Male` and `Eyeglasses` classifiers are using the same features, `Male` would continue to attend to the eye region, since these features would continue to be useful. However, when `Eyeglasses` are present, `Male` attends primarily to the mouth, not the eyes (Fig. 4). Thus, we verify that the high heatmap score `Eyeglasses` is caused by co-localized features relevant to both attributes.

Blond Hair and Wavy Hair. We choose this pair of attributes as they relate to the same regions within the image (hair) with similar MCCs (0.34 ± 0.02 and 0.37 ± 0.05), but have very different heatmap scores. Despite both referring to the hair feature, `Blond_Hair` and `Wavy_Hair` exhibit distinct attention maps. Relative to the `Male` mask score, for `Wavy_Hair` the models attend to more to the hair region, and significantly less to the eyes, nose, and mouth. This increase for hair is larger regarding `Blond_Hair`, which also has a smaller decrease in the eye region. Overall, `Blond_Hair` has a higher heatmap score of 0.72 ± 0.02, while `Wavy_Hair` is lower at 0.65 ± 0.03.

We propose that this difference is due to the presence of an (unlabeled) confounder: one of the attributes and `Male` are both correlated with the confounder, which creates an apparent relation between `Male` and the attribute. To test this, we modified the training distribution for `Blond_Hair` and `Wavy_Hair` by training models on a subsampled training set (Fig. 5). We varied the ground-truth MCC from -0.5 to -0.1 between the target attribute and `Male` by varying proportion of the 4 subgroups within the training set, keeping the overall number constant (details in Sec. E). For `Blond_Hair`, we find that there is no statistically significant change in heatmap score, with a Kendall τ value of 0.007. However, `Wavy_Hair` demonstrates a strong correlation between MCC and heatmap score ($\tau = 0.785$), with the model bias decreasing as train set bias decreases. This indicates that there might be an unlabeled confounder present in `Blond_Hair`: there is an innate quality to the features distinct from dataset labels that create bias within the model for `Blond_Hair`, rather than the simple proportion of attributes to one another in the dataset as in `Wavy_Hair`. This allows us to better understand when debiasing techniques might work: for example, methods that attempt to rebalance the dataset for `Blond_Hair` [22, 24] might struggle since the bias is not due to the presence of blond hair, but a hidden confounder.

4. Conclusion

We propose Attention-IoU, a metric for identifying and explaining spurious correlations through attention maps. With the CelebA dataset, we show that the metric and the mask and heatmap scores reveal aspects beyond dataset labels and model accuracies, recontextualizing prior analyses of CelebA. In particular, we identify ways in which different attributes are influenced by the `Male` label: attributes can be biased more or less based on labels of the sensitive attribute and can be biased in ways beyond the correlation of labels within the dataset. These insights allow us to better understand how debiasing techniques might perform on CelebA. Future investigations of the metric on other datasets and tasks can provide more insights into the nature of biases within computer vision models.

References

[1] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards Causal Benchmarking of Biasin Face Analysis Algorithms. In *Deep Learning-Based Face Analytics*, pages 327–359. Springer International Publishing, Cham, 2021. 2

[2] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63:4:1–4:15, 2019. Conference Name: IBM Journal of Research and Development. 1

[3] 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*, pages 77–91. 2018. 1

[4] Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, 56(7):1–38, 2024. 1

[5] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does Object Recognition Work for Everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 1

[6] Remi Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. Image Counterfactual Sensitivity Analysis for Detecting Unintended Bias, 2020. arXiv:1906.06439. 2

[7] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*. 2016. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[9] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women Also Snowboard: Overcoming Bias in Captioning Models. In *Computer Vision – ECCV 2018*, pages 793–811. Springer International Publishing, Cham, 2018. 1

[10] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing Error in Object Detectors. In *Computer Vision – ECCV 2012*, pages 340–353, 2012. 3, 4

[11] Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning debiased classifier with biased committee. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 18403–18415, 2024. 2

[12] Arvind Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. UDIS: Unsupervised Discovery of Bias in Deep Visual Recognition Models. In *British Machine Vision Conference (BMVC)*, 2021. 2

[13] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation, 2020. arXiv:1907.11922. 2

[14] Seongmin Lee, Judy Hoffman, Zijie J. Wang, and Duen Horng Chau. VIsCUIT: Visual Auditor for Bias in CNN Image Classifier. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21443–21451, 2022. 2

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 1, 2

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35, 2022. 1

[17] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: De-biasing Classifier from Biased Classifier. In *Advances in Neural Information Processing Systems*, pages 20673–20684. 2020. 2

[18] Dana Pessach and Erez Shmueli. A Review on Fairness in Machine Learning. *ACM Computing Surveys*, 55(3):1–44, 2023. 1

[19] Maan Qraitem, Kate Saenko, and Bryan A. Plummer. Bias Mimicking: A Simple Sampling Approach for Bias Mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20320, 2023. 2

[20] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair Attribute Classification Through Latent Space De-Biasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9301–9310, 2021. 2

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015. 2

[22] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*, 2020. 2, 4

[23] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. 1, 2

[24] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised Learning of Debiased Representations with Pseudo-Attributes. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16721–16730, 2022. 2, 4

[25] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World, 2017. arXiv:1711.08536. 1

[26] Mingxing Tan and Quoc Le. EfficientNetV2: Smaller Models and Faster Training. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10096–10106. 2021. 4

[27] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7, 2018. 1, 2

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2

[29] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. *International Journal of Computer Vision*, 130(7):1790–1810, 2022. 1

[30] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating Bias and Fairness in Facial Expression Recognition. In *Computer Vision – ECCV 2020 Workshops*, pages 506–523, 2020. 2

[31] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, pages 818–833, 2014. 1

[32] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and Evaluating Racial Biases in Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14830–14840, 2021. 1

[33] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017. 1, 2

[34] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 1

[35] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. 2

Attention IoU: Examining Biases in CelebA using Attention Maps

Supplementary Material

A. Gradients for GradCAM

In Sec. 2, we use GradCAM to calculate attention maps. Given an input image $\mathbf{x}$ and target attribute a, GradCAM computes the gradient of the class output y_a with respect to the output of a convolutional layer, usually the final layer, to obtain activation maps of the attribute. A simple gradient-weighted linear combination of the layer's feature activation maps produces the attribute-specific attention map $\mathrm{GradCAM}_a(\mathbf{x})$. GradCAM was developed for models trained with categorical cross entropy loss, and thus, in its standard implementation, only able create attention maps for positive predictions for a model trained with binary cross entropy loss. For our metric, we instead take the gradient of the absolute value of the class output, $|y_a|$, so that image features that contribute positively to either prediction is attended to in the attention map.

When using a model that is trained using binary cross-entropy loss, computing the gradient w.r.t. the absolute value of the logit (before the sigmoid) is equivalent to computing the gradient w.r.t. to the predicted class for categorical cross-entropy loss with two heads (one each for the positive and negative class). Concretely, let s be the value of the logit; the probability that this model assigns to the positive class is $\sigma(s) = \frac{1}{1+e^{-s}}$, and the probability assigned to the negative class is $1 - \sigma(s) = \frac{e^{-s}}{1+e^{-s}} = \sigma(-s)$. The model prediction is $\arg\max(\sigma(s), \sigma(-s)) = \arg\max(s, -s)$. Thus, taking the gradient with respect to the absolute value of the logits allows us to find positive contributions to the predicted binary class.

B. Proofs of Invariants

In Sec. 2, we introduce the Attention-IoU metric, $\mathcal{B}_{\text{A-IoU}}$, which is invariant to scale and size for pixel maps.

First, we confirm that if the two input maps are identical, $\mathbf{M}_1 = \mathbf{M}_2 = \mathbf{M} \in \mathbb{R}^{h \times w}$, the Attention-IoU metric is 1:

$$\mathcal{B}_{\text{A-IoU}}(\mathbf{M}, \mathbf{M}) = \frac{\langle \widehat{\mathbf{M}}, \widehat{\mathbf{M}} \rangle_F}{\left\| \frac{\widehat{\mathbf{M}} + \widehat{\mathbf{M}}}{2} \right\|_F^2} \tag{1}$$

$$= \frac{\langle \widehat{\mathbf{M}}, \widehat{\mathbf{M}} \rangle_F}{\left\| \widehat{\mathbf{M}} \right\|_F^2} = \frac{\left\| \widehat{\mathbf{M}} \right\|_F^2}{\left\| \widehat{\mathbf{M}} \right\|_F^2} = 1. \tag{2}$$

We next prove that $\mathcal{B}_{\text{A-IoU}}$ is scale invariant. Given two maps $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{h \times w}$, suppose the maps are multiplied by the scalars $a_1, a_2 \in \mathbb{R}_+$ respectively. Then their L_1 normalized maps are

$$\widehat{a_i \mathbf{M}_i} = \frac{a_i \mathbf{M}_i}{||a_i \mathbf{M}_i||_1} = \frac{a_i \mathbf{M}_i}{a_i ||\mathbf{M}_i||_1} = \widehat{\mathbf{M}}_i \tag{3}$$

So $\mathcal{B}_{\text{A-IoU}}(a_1 \mathbf{M}_1, a_2 \mathbf{M}_2) = \mathcal{B}_{\text{A-IoU}}(\mathbf{M}_1, \mathbf{M}_2)$.

For the proof of size invariance, we assume for simplicity that the maps are resized by a positive integer scalar $\alpha \in \mathbb{N}$ using nearest neighbor interpolation. Again, consider two maps $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{h \times w}$. Let $\mathbf{M}_1^\alpha, \mathbf{M}_2^\alpha \in \mathbb{R}^{\alpha h \times \alpha w}$ be the rescaling of the two maps by the constant α. For example, with $\alpha = 2$, a 5×5 box in the center of the map will be resized to be a 10×10 box, with the same spacial location within the map. Note that the L_1 normalized maps are

$$\widehat{\mathbf{M}}_i^\alpha = \frac{\mathbf{M}_i^\alpha}{||\mathbf{M}_i^\alpha||_1} = \frac{\mathbf{M}_i^\alpha}{\alpha^2 ||\mathbf{M}_i||_1}, \tag{4}$$

as each pixel in the original map appears α^2 times in the resized map. Furthermore, the Frobenius inner product of the two resized maps is

$$\langle \mathbf{M}_1^\alpha, \mathbf{M}_2^\alpha \rangle_F = \sum_{i=1}^{\alpha h} \sum_{j=1}^{\alpha w} (\mathbf{M}_1^\alpha)_{ij} \cdot (\mathbf{M}_2^\alpha)_{ij} \tag{5}$$

$$= \alpha^2 \sum_{i=1}^{h} \sum_{j=1}^{w} (\mathbf{M}_1)_{ij} \cdot (\mathbf{M}_2)_{ij} \tag{6}$$

$$= \alpha^2 \langle \mathbf{M}_1, \mathbf{M}_2 \rangle_F \tag{7}$$

and, for the norm,

$$\left\| \frac{\widehat{\mathbf{M}}_1^\alpha + \widehat{\mathbf{M}}_2^\alpha}{2} \right\|_F^2 = \frac{1}{4} \sum_{i=1}^{\alpha h} \sum_{j=1}^{\alpha w} \left(\frac{(\mathbf{M}_1^\alpha)_{ij}}{||\mathbf{M}_1^\alpha||_1} + \frac{(\mathbf{M}_2^\alpha)_{ij}}{||\mathbf{M}_2^\alpha||_1} \right)^2 \tag{8}$$

$$= \frac{1}{4\alpha^4} \sum_{i=1}^{\alpha h} \sum_{j=1}^{\alpha w} \left(\frac{(\mathbf{M}_1^\alpha)_{ij}}{||\mathbf{M}_1||_1} + \frac{(\mathbf{M}_2^\alpha)_{ij}}{||\mathbf{M}_2||_1} \right)^2 \tag{9}$$

$$= \frac{1}{4\alpha^2} \sum_{i=1}^{h} \sum_{j=1}^{w} \left(\frac{(\mathbf{M}_1)_{ij}}{||\mathbf{M}_1||_1} + \frac{(\mathbf{M}_2)_{ij}}{||\mathbf{M}_2||_1} \right)^2 \tag{10}$$

$$= \frac{1}{\alpha^2} \left\| \frac{\widehat{\mathbf{M}}_1 + \widehat{\mathbf{M}}_2}{2} \right\|_F^2. \tag{11}$$

Thus, combining the two parts together,

$$\mathcal{B}_{\text{A-IoU}}(\mathbf{M}_1^\alpha, \mathbf{M}_2^\alpha) = \frac{\langle \widehat{\mathbf{M}}_1^\alpha, \widehat{\mathbf{M}}_2^\alpha \rangle_F}{\left\| \frac{\widehat{\mathbf{M}}_1^\alpha + \widehat{\mathbf{M}}_2^\alpha}{2} \right\|_F^2} \tag{12}$$

$$= \frac{\frac{1}{\alpha^4 ||\mathbf{M}_1||_1 \cdot ||\mathbf{M}_2||_1} \langle \mathbf{M}_1^\alpha, \mathbf{M}_2^\alpha \rangle_F}{\frac{1}{\alpha^2} \left\| \frac{\widehat{\mathbf{M}}_1 + \widehat{\mathbf{M}}_2}{2} \right\|_F^2} \tag{13}$$

$$= \frac{\frac{1}{\alpha^2} \frac{1}{||\mathbf{M}_1||_1 \cdot ||\mathbf{M}_2||_1} \langle \mathbf{M}_1, \mathbf{M}_2 \rangle_F}{\frac{1}{\alpha^2} \left\| \frac{\widehat{\mathbf{M}}_1 + \widehat{\mathbf{M}}_2}{2} \right\|_F^2} \tag{14}$$

$$= \frac{\langle \widehat{\mathbf{M}}_1, \widehat{\mathbf{M}}_2 \rangle_F}{\left\| \frac{\widehat{\mathbf{M}}_1 + \widehat{\mathbf{M}}_2}{2} \right\|_F^2} \tag{15}$$

$$= \mathcal{B}_{\text{A-IoU}}(\mathbf{M}_1, \mathbf{M}_2). \tag{16}$$

Although in the proof $\mathbf{M}_1^\alpha$ and $\mathbf{M}_2^\alpha$ are larger matrices than $\mathbf{M}_1$ and $\mathbf{M}_2$, the same argument applies if $\mathbf{M}_1$ and $\mathbf{M}_2$ are zero-padded to have same dimensions as the resized maps.

C. Experimental Setup

Since we require ground-truth segmentation masks, we use CelebAMask-HQ [13] which is a subset of 30,000 images from CelebA [15], in which each image has a high-quality segmentation mask of different facial features, including hair, nose, skin, hats, and jewelry. We group like features together, *e.g.*, {left brow, right brow} and {upper lip, lower lip, mouth}. Large non-localized feature masks (background, skin, and cloth) are excluded from our analyses. We choose a 70%-15%-15% train-validation-test split for training on CelebAMask-HQ. The test set was used to compute the overall accuracy, per-group accuracy, and Attention-IoU. To train classifiers for the attributes, we use a ResNet-50 model [8] pretrained on ImageNet [21]. We replaced the final layer with two fully-connected layers with a hidden layer size of 2,048 and a dropout layer between them, in order to improve accuracy, following Ramaswamy *et al*. for their CelebA ResNet classifier [20]. Input images are rescaled to be 224×224, and augmented using random crops and horizontal flips during training. We used a binary cross-entropy loss, weighted proportionally to positive examples of each attribute. Models were trained for 10 epochs, with a batch size of 32. We report averages and standard deviations over 20 individually trained models.

D. Validating the metric

In addition to evaluating on CelebA, we test the proposed metric on Waterbirds [22]. This simple synthetic dataset is constructed by combining cropped bird images from the CUB dataset [28] with backgrounds from the Places dataset [35]. In the dataset birds are labeled as either a waterbird or landbird, and backgrounds are similarly labeled as land or water. The dataset can be constructed with different levels of correlation between the bird and the background, introducing a single axis of bias within the dataset. Moreover, masks of the bird and background are clearly available within this dataset, which can be used to compute Attention-IoU.

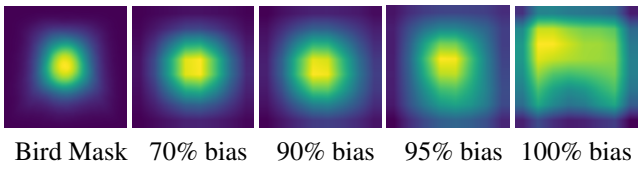


Figure 6. **Average bird mask and average heatmaps for Waterbirds at increasing levels of bias**. We see that the model attends less on the bird as the bias increases, as indicated by its mask.

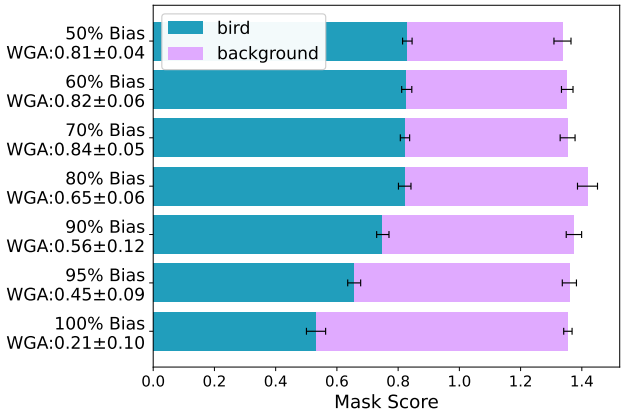


Figure 7. **Evaluation of mask score using GradCAM on Waterbirds test set**. The X-axis represents the Attention-IoU mask score for the ground-truth masks of the bird and background. We note the dataset bias and the worst group accuracy (**WGA**) along the Y-axis. As the bias increases, the worst group accuracy decreases and the model attends less to the bird and more to the background.

Experimental setup. Following prior work, we place a specified percentage (between 50%-100%) of the waterbirds on a water background, with the remaining 0%-50% of the waterbirds are placed on a land background, and similarly for landbirds and land backgrounds. The validation and test sets are unbiased with a bird being 50% likely to align with its background. We followed Sagawa *et al*. [22] in using the official train-test split of the CUB dataset, composed of 5,994 training images and 5,794 testing images, and randomly choosing 20% of the training images to form the validation set. As our model, we used ResNet-18 [8] pretrained on ImageNet [21]. Models were trained on Waterbirds using categorical cross-entropy loss with a batch size of 64. Other hyperparameters remain the same as Sec. C.

Results. We compare the heatmap generated with the ground-truth masks for the bird. In Fig. 6, we show the average bird mask, as well as the average heatmaps generated by GradCAM across all images in the test set for models trained at different levels of bias. As the bias increases, models rely more on cues from the background. This is reflected in the heatmaps, which highlight regions other than the bird mask. We verify that Attention-IoU captures this effect in Fig. 7, which shows the mask scores across varying training set bias for both bird and background masks. We also report the worst group accuracy (WGA) of models for each. As expected, the worst group accuracy decreases from 0.81 ± 0.02 to 0.21 ± 0.10 as bias increases from 50% to 100%. The decrease from 0.72 ± 0.02 to 0.42 ± 0.03 in mask score almost exactly mirrors the proportional decrease in WGA, validating that the metric accurately measures model bias. Due to the simple nature of Waterbirds, the bias in the dataset is directly represented in the training distribution, and Attention-IoU captures this perfectly.

E. Subsampling Training Details

Here we provide experimental details for varying training set correlations in Sec. 3. Given a target Matthews correlation coefficient between the specified attribute and `Male`, we find subgroup sizes that achieve the target MCC (as MCC is dependent entirely on the sizes of the 4 subgroups) using SciPy's `optimize.minimize` with the trust region method[2] (Fig. 8). We bound the sizes of the subsampled subgroups to the size of the original groups, and aim to minimize the distance to the original group sizes by the L_2 norm. To reduce fluctuations between the subsampled sizes, we initialize the optimizer with the adjacent subgroup sizes, with the original subgroups sizes in the training set as the starting point. Lastly, after running the optimization once for all MCCs, we rerun the optimization process with the additional bound of the smallest subsampled training set, so that all the subsampled training sets are of the same size. As the subsampling was an ablation study, the heatmap scores reported in Fig. 5 were run on the validation set.

F. Additional CelebA Results

Model Evaluation. The average precision weighted for all 40 attributes in CelebA, averaged across the 20 trained models with the experimental setup detailed in Sec. C, is 0.902 ± 0.025. For reference, the normalized average precision (AP_N) [10] for the `Male` attribute is 0.994 ± 0.003, the second highest after `Eyeglasses` (0.998 ± 0.001). In Fig. 9 we show average heatmaps for select attributes.

CelebA Normalized Average Precision. As a comparison to Fig. 2, which shows CelebA mask score against worst

[2] https://docs.scipy.org/doc/scipy/reference/optimize.minimize-trustconstr.html

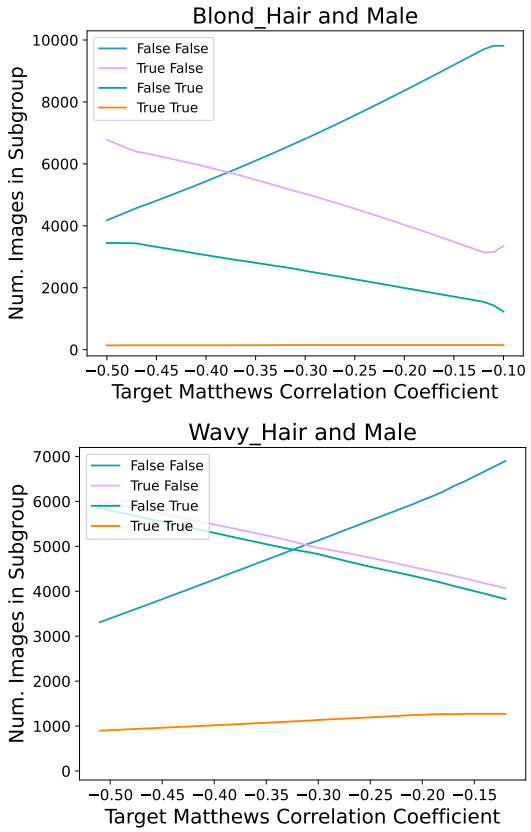


Figure 8. **Training set subgroup sizes under subsampling**. Here we report subgroup sizes of the training set of varying MCCs for `Blond_Hair` and `Wavy_Hair` with `Male`, under our optimization scheme, to compute the results in Fig. 5. Subgroup sizes are bounded to the smallest subsampled training set size. The legend shows the four different subgroups groups, with the first value indicating the target label and the second `Male`.

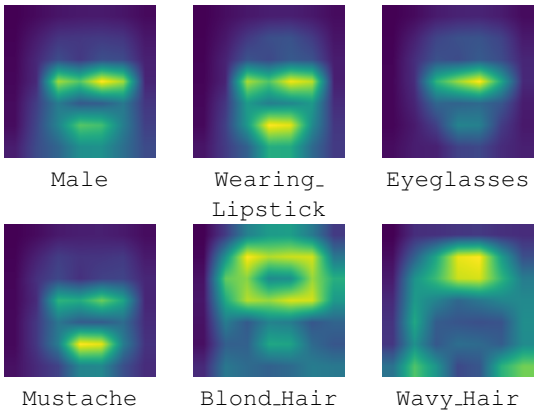


Figure 9. **Average heatmaps for CelebA attributes**. We visualize average heatmaps for the selected attributes investigated in Sec. 3.

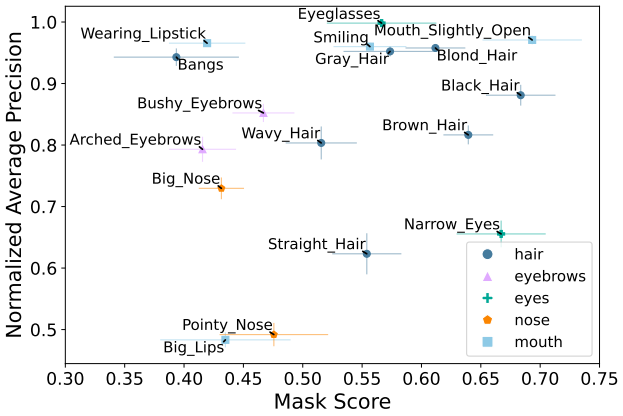
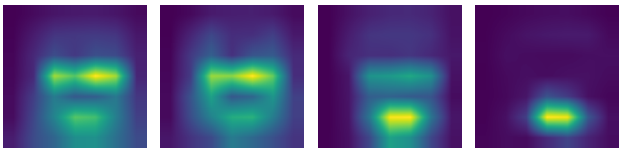


Figure 10. **Evaluation of mask score using GradCAM on CelebA test set with attribute-specific feature masks, compared to average precision**. To compare per-attribute AP between attributes, we adopt Hoiem *et al.*'s normalized average precision (AP_N) metric [10].

group accuracy, in Fig. 10 we show the mask score of the same 17 attributes to their normalized average precision (AP_N). Compared with worst group accuracy, there is a no correlation for normalized average precision with respect to the mask score. Unlike worst group accuracy, to calculate normalized average precision one does not need to assume the correlated attribute.

Mustache. In addition to the four attributes analyzed in Sec. 3, we also analyze `Mustache` as another example of an outlier to the heatmap score trend (Fig. 3 *left*). `Mustache` is moderately correlated with `Male`, with a predicted label MCC of 0.51 ± 0.04. `Mustache`'s mask score distribution reflects that of `Male`, with slightly more attention to the hair and mouth regions. This is reflected by a high heatmap score of 0.90 ± 0.02. We choose this attribute since this attribute represents a one-way correlation: images where `Mustache` are labeled as present are almost often labeled `Male`, whereas images where `Mustache` are labeled as absent are roughly evenly split among being labeled `Male` and not `Male`.

We investigate how Attention-IoU changes based on the ground-truth values of these attributes (Fig. 11). The score is extremely high (0.94 ± 0.02) among images labeled not `Male`. When `Male` is false, the `Mustache` and `Male` attention maps closely align, indicating that the model is heavily relying on `Male` to classify `Mustache`. However, when the image is labeled as `Male`, the score is lower (0.84 ± 0.5 and 0.82 ± 0.03 for `Mustache` true and false respectively), the models attend less to `Male` regions in order to classify `Mustache`. `Mustache` demonstrates that even though two attributes may be one-way predictive in the dataset (and thus have a lower MCC), the models still attend strongly to any correlation between the attributes, which is indicated through Attention-IoU.



`Male` heatmap | not `Male`, no `Mustache` | `Male`, no `Mustache` | `Male`, `Mustache`

Figure 11. **Average heatmaps for `Mustache`**. We visualize average heatmaps for `Mustache` for images where `Mustache` and `Male` are labeled false (*center-left*), where `Mustache` is labeled false and `Male` is labeled true (*center-right*) and where `Mustache` and `Male` are labeled true (*far right*), and compare to the `Male` heatmap (*far left*). When `Male` is labeled as false, `Mustache` and `Male` attention maps closely align but do not when `Male` is labeled true.

G. Evaluating with EfficientNet

To demonstrate the effectiveness of Attention-IoU on architectures other than ResNet, we also evaluated the metric using the EfficientNetV2-S architecture [26] on both the Waterbirds and CelebA datasets. Aside from the change in architecture, and averaging over 10 trained models instead of 20, the experimental setup remained the same.

For Waterbirds, the EfficientNet models show a very similar pattern to ResNet in attending less to the bird and more to the background as dataset bias increases (Fig. 13). The EfficientNet heatmap scores for CelebA also show a strong positive trend with MCC like ResNet (Fig. 12). The 5 highlighted attributes maintain their relative positions, with some changes owing to different architectures and pretraining weights.

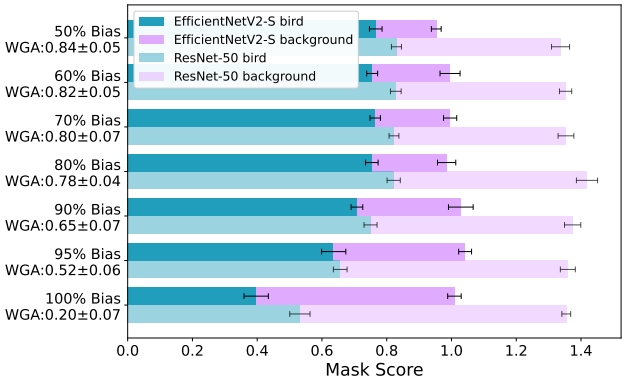


Figure 12. **EfficientNetV2 mask score on Waterbirds**. The top bars indicate Attention-IoU mask scores for EfficientNetV2-S models, while the bottom bars are corresponding ResNet-50 scores from Fig. 3. WGA is for the EfficientNet model. As with ResNet, the EfficientNet models attend less to the bird and more to the background, mirroring the decrease in WGA.

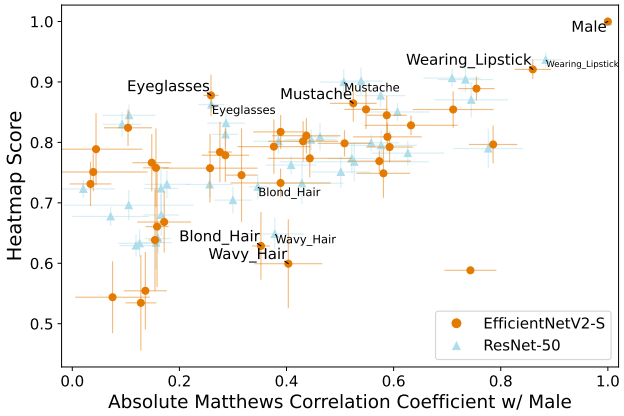


Figure 13. **EfficientNetV2 heatmap scores on CelebA attributes**. Orange indicates results with EfficientNetV2-S models, and light blue are ResNet-50 results from Fig. 5. We observe a very similar trend in EfficientNetV2 to that of ResNet-50. Highlighted attributes maintain their relative position, with some movement owing to different architectures and pretraining weights.