A Multidimensional Analysis of Social Biases in Vision Transformers

Jannik Brinkmann*   Paul Swoboda   Christian Bartelt

University of Mannheim

{jannik.brinkmann, paul.swoboda, christian.bartelt}@uni-mannheim.de

Abstract

The embedding spaces of image models have been shown to encode a range of social biases such as racism and sexism. Here, we investigate specific factors that contribute to the emergence of these biases in Vision Transformers (ViT). Therefore, we measure the impact of training data, model architecture, and training objectives on social biases in the learned representations of ViTs. Our findings indicate that counterfactual augmentation training using diffusion-based image editing can mitigate biases, but does not eliminate them. Moreover, we find that larger models are less biased than smaller models, and that models trained using discriminative objectives are less biased than those trained using generative objectives. In addition, we observe inconsistencies in the learned social biases. To our surprise, ViTs can exhibit opposite biases when trained on the same data set using different self-supervised objectives. Our findings give insights into the factors that contribute to the emergence of social biases and suggests that we could achieve substantial fairness improvements based on model design choices.

1. Introduction

In recent studies, state-of-the-art self-supervised image models such as SimCLR [9] and iGPT [8] have been shown to encode a range of social biases, such as racism and sexism [34]. This can lead to representational harm [4] and ethical concerns in different socio-technical application scenarios [41]. The distributional nature of these models is suspected to be an important factor contributing to the emergence of social biases, as it has been demonstrated that these models tend to encode common co-occurrences of objects associated with social biases (e. g. women are more often set in "home or hotel" scenes, whereas men are more often depicted in "industrial and construction" scenes [36]). Moreover, it has been demonstrated that self-supervised training objectives can impact the distribution of social biases in models that share the same ResNet50 [14] architecture [33].

---

* Corresponding author. The associated code is available at github.com/jannik-brinkmann/social-biases-in-vision-transformers.
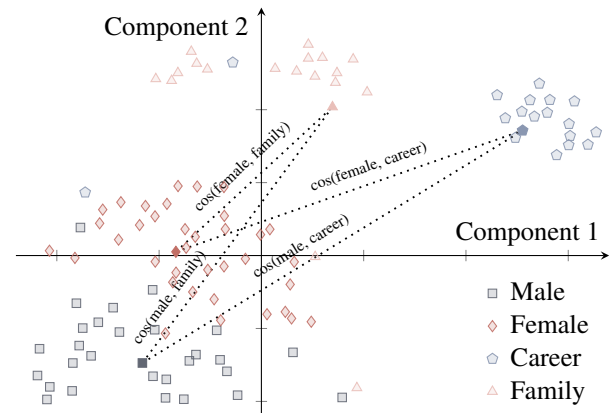

Figure 1: Gender bias in image embedding from ViTMAE: t-SNE (n=2) reveals that "female" is more closely associated with "family" rather than "career", whereas "male" has a comparable association with both attributes.

However, existing work has done little investigation into other factors that contribute to the emergence of social biases in image models.

Contributions   Here, we seek to better understand the factors that contribute to the emergence of social biases in image models. Therefore, we investigate social biases in embedding spaces, which, despite not being observable for end-users, could propagate into downstream tasks during fine-tuning. This can help to make informed choices about the model to select for a downstream task, and to develop effective strategies to mitigate social biases. In detail, the contributions of our work are:

- Training ViTs with counterfactual data augmentation using diffusion-based image editing can reduce social biases, but is not sufficient to eliminate them.

- ViTs trained using discriminative objectives are less biased than those trained using generative objectives.

- Scaling ViTs can help to mitigate social biases.

- ViTs can exhibit opposite biases despite being trained on the same data set, which indicates that biases are not just a result of simple object co-occurrences.
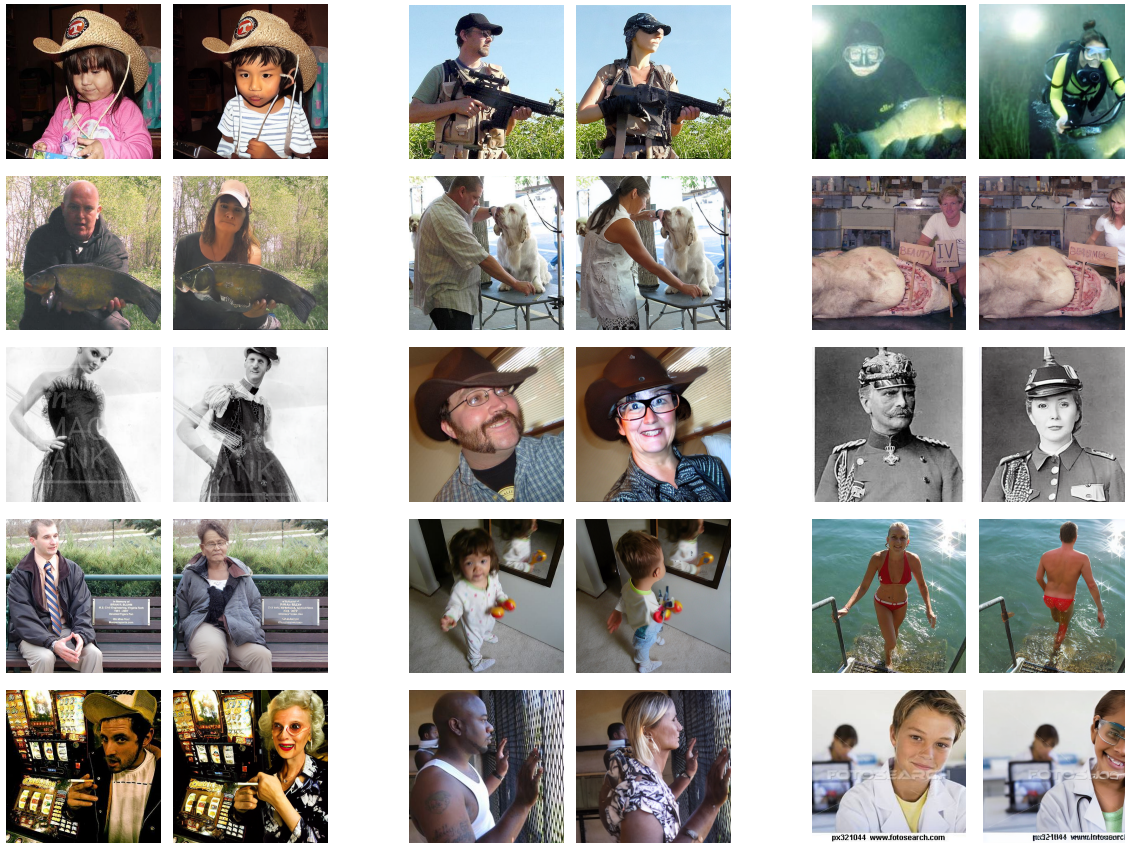
Figure 2: **Selected counterfactual images on ImageNet.** In each case, we show the original image (left), and the generated counterfactual image (right).

2. Related Work

Self-Supervised Learning of ViTs Self-supervised approaches have emerged as the standard for training large machine learning models since they don't require labeled data and learn representations that generalize well across different downstream tasks [6]. Transformer models [35], which were designed as sequence-to-sequence models for natural language translation, have been adapted to computer vision [12]. Self-supervised learning techniques applied to ViTs can be classified into discriminative (or joint-embedding) methods and generative (or reconstruction-based) methods [32]. Discriminative methods encourage similarity among representations from diverse augmentations of a given input image, while generative methods utilize a reconstruction loss that does not rely on augmentations. Instead it uses a decoder to reconstruct the original image given a masked image. Both methods have demonstrated strong empirical results on downstream tasks [7, 8, 10, 13].

Social Biases in Image Embeddings The embeddings of self-supervised image models have been shown to encode a range of human-like social biases [34]. However, the analysis was confined to SimCLR [9] and iGPT [8] as embedding models. Therefore, Sirotkin *et al.* [33] built on this work to examine the distribution of social biases in image models that were trained using a range of self-supervised objectives, such as geometric, cluster-based, and contrastive methods. The authors discovered that models trained with contrastive methods exhibit the largest number of social biases, and that the distribution of biases differs depending on the studied embedding layer. However, their analysis focused only on training objectives and the number of social biases without considering the direction of the bias, constraining the interpretability of their findings. In addition, their investigation was conducted on models using a ResNet50 [14] architecture, excluding ViTs which are considered the standard for transfer learning [17].

Bias Mitigation Methods The approaches to mitigate biases can be distinguished into methods that manipulate the training data and methods that adjust the training procedure [23]. To mitigate biases during training, existing work suggests, amongst others, adversarial learning [37], train-

ing separate models for each attribute [38], or incorporating regularization terms [2, 15]. In contrast, the methods to mitigate biases in the training data aim to generate unbiased data sets that are balanced [16] or do not include information about the bias dimension [24]. One approach to mitigate biases in the training data is Counterfactual Data Augmentation (CDA) [45]. This method entails generating training instances that contradict the observed biases. There are different variations of CDA: 1-sided CDA, which use just the counterfactuals during an additional pre-training phase, and 2-sided CDA, which uses both counterfactuals and the original training data. While 1-sided CDA has a more substantial impact on biases, it can lead to over-correction [39]. In existing work, CDA has been used to mitigate different types of biases in language models [20], operating on a set of term pairs, such as "man" and "woman". However, generating counterfactual training instances from images is non-trivial. To address this, conditional generative adversarial networks have been used to generate unbiased training data with balanced protected attributes [26, 31]. Therefore, the authors generate multiple synthetic images for each training image, maintaining the target attribute score but reversing the expression score on the protected attribute. These approaches have demonstrated to be effective at mitigating bias on selected dimensions, but do not eliminate them. In addition, existing methods focus on downstream tasks and no research has been conducted on debiasing pre-trained image models used as backbones for transfer learning.

3. Background

iEAT The Image Embedding Association Test (iEAT) quantifies social biases in image embeddings based on semantic similarities [34]. It compares the differential association of image embeddings of selected target concepts (such as "male" and "female") and attributes (such as "science" and "liberal arts"), and tests the null-hypothesis of equal similarities of the target concepts and attributes. Hence, a rejection suggests that one target concept is more associated with one attribute than the other (such as "male" is more associated with "science" or "female" is more associated with "liberal arts"). To test the null-hypothesis, it formulates a test statistic that compares target concepts X and Y with attributes A and B, defined as:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where $s(w, A, B)$ is the differential association of a target concept with the attributes, measured using the cosine similarities of their embeddings:

$$s(w, A, B) = \mu(cos(w, a)_{a \in A}) - \mu(cos(w, b)_{b \in B})$$

where μ is the mean. The statistical significance is determined using a permutation test, contrasting the score $s(X, Y, A, B)$ with the scores $s(X_i, Y_i, A, B)$, where X_i and Y_i are all equal-sized partitions of the set $X \cup Y$:

$$p_t = Pr[s(X_i, Y_i, A, B) > s(X, Y, A, B)] \qquad (1)$$

The effect size d quantifies the bias magnitude, computed as the normalized separation of the association distributions:

$$d = \frac{\mu(s(x, A, B)_{x \in X}) - \mu(s(y, A, B)_{y \in Y})}{\sigma(s(t, A, B)_{t \in X \cup Y})} \qquad (2)$$

where μ is the mean and σ is the standard deviation. Here, the distance from zero indicates the bias magnitude, such that an effect size equaling zero implies the absence of bias. Moreover, the effect size indicates the direction of the bias, such that a negative effect size suggests that the differential association of Y with A and B is more pronounced, whereas a positive effect size implies the opposite scenario.

The iEAT framework introduces a collection of 15 association tests designed to measure human-like social biases (see Table 1). These tests offer a valuable baseline to assess the presence and intensity of certain social biases within image embeddings. However, it it important to recognize that these are not an exhaustive list of all possible biases. These biases were selected due to their recurrence in related literature and societal implications. However, there might be other biases not captured in this selection, such as political biases. Nonetheless, these tests remain an instrumental foundation to assess the existence and magnitude of social biases in image embeddings.

Embedding Layer The selection of an embedding layer is crucial to extract features that contain high-quality, general-purpose information about the objects in an image. It has been demonstrated that in ViTs trained with supervised methods, the model depth tends to correlate with the quality of the embeddings, with the highest-quality embeddings being in the second-to-last layer [42]. In contrast, ViTs trained with SSL methods have been found to generate the most useful embeddings at a layer in the middle of the model [3, 8]. Therefore, the selection of an embedding layer depends on the training approach and the specific model. Here, for each model, we choose the layer that has been reported to be optimal in linear evaluations.

4. Experiments and Results

Here, we describe and discuss our experiments to investigate factors that contribute to the emergence of social biases in the embedding spaces of ViTs. Therefore, we assess bias mitigation methods along multiple dimensions:

- Training data: We investigate counterfactual augmentation training using diffusion-based image editing and find

that it can reduce social biases in ViTs, but is not sufficient to eliminate them (Section 4.1).

- Training objectives: We assess the impact of training objectives, and find that ViTs trained using discriminative objectives are less biased than those trained using generative objectives (Section 4.2).

- Model architecture: We evaluate the impact of different architectural choices and find that social biases decrease as model size and input resolution increase, but observe no systematic effect for patch size (Section 4.3).

4.1. Impact of Training Data

The emergence of social biases in self-supervised image models is often suggested to be a result of object co-occurrences in images (*e.g.* women are more often set in "home or hotel" scenes, whereas men are more often depicted in "industrial and construction" scenes [36]). However, little research has been conducted on the effect of modifications of the training data on social biases in pre-trained image models. Therefore, we investigate the debiasing effect of counterfactual data on gender bias as an example. Our findings suggest that it can reduce social biases both during pre-training and fine-tuning, although it does not eliminate them and can come at a cost of a slight reduction in downstream performance. Moreover, we observe differences in the responsiveness to the counterfactual data, suggesting that its effectiveness is model-specific.

Models In our experiments, we use BEiT [3], ViT-MoCo [10] and ViT-MAE [13], which use a standard Transformer as the backbone network (12 layers, 12 attention heads, 768 hidden size). The implementation and model weights were made available using HuggingFace's Transformers [35] and Timm [40].

Counterfactual Data Augmentation To investigate the impact of training data, we examine to what extent counterfactual data augmentation can mitigate social biases in ViTs. In our experiments, we combine the approach to counterfactual data augmentation used in natural language processing with diffusion-based image editing. Therefore, we leverage a large-scale text-to-image diffusion model [28] as a foundation, to capitalize on the benefits of pre-training on a sizable and generic corpus. For each image, we generate a textual description using BLIP [21] and CLIP [25]. Then, we use a set of term pairs (*e.g.* "man", "woman") to substitute target words in the generated caption. For our purposes, we adopt the set of gender term pairs of Zhao *et al.* [44]. To generate counterfactual images, we use diffusion-based semantic image editing with mask guidance [11]. To this end, we use CLIPSeg [22] to mask the target words (*e.g.* "man") in the image and use Stable Diffusion [29] to inpaint the masked image section, conditioned on the modified captions (see Figure 2).

Here, we adopt the ImageNet ILSVRC 2012 dataset (ImageNet-1K) [30] as our benchmark to assess the effectiveness of the generated data, as it is one of the most studied benchmarks for which there is an extensive literature on architecture and training procedures. ImageNet-1K contains 1.28 million images, from which we generate an additional 159,393 counterfactual images.

Counterfactual Training To evaluate the debiasing effect of counterfactual data, we follow Webster *et al.* [39] and continue the training of the models from a pre-trained checkpoint using the counterfactual images (1-sided CDA). To this end, we adopt the standard contrastive learning objective for ViT-MoCo [10] and masked image modeling training objective for BEiT and ViT-MAE with a masking ratio of 40 % [3] and 75 % [13], respectively. Then, we train each model using Adam [18] with a batch size of 128

Test	Target A	Target B	Attribute X	Attribute Y
T1	Young	Old	Pleasant	Unpleasant
T2	Other	Arab-Muslim	Pleasant	Unpleasant
T3	European American	Asian American	American	Foreign
T4	Disabled	Not-Disabled	Pleasant	Unpleasant
T5	Male	Female	Career	Family
T6	Male	Female	Science	Liberal Arts
T7	Flower	Insect	Pleasant	Unpleasant
T8	European American	Native American	Pleasant	Unpleasant
T9	European American	African American	Pleasant	Unpleasant
T10	Christianity	Judaism	Pleasant	Unpleasant
T11	Gay	Straight	Pleasant	Unpleasant
T12	Light Skin	Dark Skin	Pleasant	Unpleasant
T13	White	Black	Tool	Weapon
T14	White	Black	Tool	Weapon (Modern)
T15	Thin	Fat	Pleasant	Unpleasant

Table 1: Image Embedding Association Tests

and learning rate 1.5e-4 for a single epoch to avoid over-correction [39]. The results are depicted in Table 2.

Model	Baseline		CDA	
	Bias	Cifar10	Bias	Cifar10
BEiT	0.65	**87.5**	**0.45**	84.8
ViT-MoCo	1.41	**95.1**	**1.39**	**95.1**
ViT-MAE	**0.59**	89.6	0.64	**89.6**

Table 2: iEAT effect size (see Equation 2) and linear evaluation performance on CIFAR10 of different models before (Baseline) and after (CDA) debiasing using a single pre-training epoch on counterfactual data. **We find that counterfactual data augmentation can reduce social biases, but its effect is model-specific and can come with a reduction in representation quality.**

In addition to the gender bias, we report the linear evaluation performance on CIFAR10 [19] as a measure of representation quality. We observe that it does reduce gender bias on BEiT and ViT-MoCo but comes with a slight reduction in representation quality for BEiT. However, a similar effect has been observed in alternative debiasing methods before and is not specific to out setting [27, 43]. In contrast, we observe the opposite effect on ViT-MAE, where it comes with a small increase in gender bias. This implies that there are differences in the responsiveness to the counterfactual data, suggesting that the effectiveness of this technique might be model-specific. We hypothesis that this is a result of the training objectives, which could influence how the models learn from the counterfactual data. In addition, we conjecture that the counterfactual data could interact differently with pre-trained checkpoints, which could carry certain biases leading to varying debiasing effects.

Model	Baseline		CDA	
	Bias	Cifar10	Bias	Cifar10
ViT-MoCo	1.25	90.4	**1.04**	**90.9**
ViT-MAE	**0.50**	**82.9**	0.55	71.2

Table 3: iEAT effect size (see Equation 2) and linear evaluation performance on CIFAR10 of different models pre-trained from scratch on ImageNet-1k (Baseline), and both ImageNet-1k and the counterfactual data (CDA). **We again observe a decrease in gender bias on MoCo-v3 and increase on ViTMAE. This implies that the observed effects are not a result of the pre-trained checkpoint.**

To evaluate whether the observed effects on ViT-MoCo and ViT-MAE are a result of their pre-trained checkpoints, we train them from scratch on ImageNet-1k and our counterfactual data (2-sided CDA). The results are illustrated in

Table 3. We again observe a decrease in gender bias on ViT-MoCo, and a similar increase in gender bias on ViT-MAE. This implies that observed effects are not a result of the pre-trained checkpoint, and that other factors influence the debiasing effect, such as model architecture differences. These findings highlight the nuanced effect of training data on social biases, demanding tailored approaches for different architectures and training approaches. Thus, we anticipate the need for more principled approaches that eliminate undesirable model behavior, potentially bypassing the use of counterfactual data and instead using post-hoc interventions to eliminate biases directly.

4.2. Impact of Training Objectives

ResNet50 [14] models, when trained using different self-supervised objectives exhibit a different number of social biases [33]. Therefore, we investigate the effect of training objectives on biases in ViTs across a range of different self-supervised methodologies: discriminative and generative models. Our findings indicate that ViTs trained with discriminative learning objective are less biased than those trained using generative objectives. Moreover, we observe that models trained on the same dataset using different objectives can exhibit opposite biases, which highlights the importance of training objectives as an important factor in the emergence of social biases in embedding spaces.

Discriminative and Generative Objectives We investigate the distribution of social biases in ViTs trained on ImageNet-21k using different self-supervised objectives. To this end, we follow Sirotkin *et al.* [33] and count the number of significant social biases across different values of p_t (see Equation 1) in the range of $[10^{-4}, 10^{-1}]$, where lower values of p_t correspond to higher statistical significance of the social biases. The results of this analysis are illustrated in Figure 3. Our findings indicate that, on average, ViTs trained using discriminative objectives exhibit fewer biases than those trained using generative objectives. This effect remains consistent across all threshold values, which highlights the robustness of our findings. We conjecture that this stems from the inherent characteristics of models trained using generative objectives, which encourage the model to reconstruct images that match the statistical patterns in the training data, capture underlying structure and dependencies within the data. Thus, if the training data is biased towards specific demographics, objects, or scenes, the model could unintentionally learn and perpetuate those biases in its representations. In contrast, discriminative learning objectives encourage representations that maximize view invariance between samples from the same image [32]. This encourages the model to learn and prioritize fundamental visual features that are less influenced by social biases or external factors.

MODELS	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
DISCRIMINATIVE MODELS															
VIT-DINO-B	**0.99**	**1.20**	−0.86	0.88	**0.38**	0.01	−0.12	**0.84**	0.49	0.22	−0.08	−0.13	−0.88	−0.77	**1.24**
VIT-MOCO-B	−0.15	**1.02**	−0.75	−0.29	**1.41**	0.13	**1.68**	−0.66	**1.10**	0.46	−0.24	−0.11	0.77	0.14	0.64
VIT-MSN-B [1]	0.93	**1.24**	0.33	0.93	0.14	−0.31	0.10	0.60	−0.78	0.54	−0.28	**−1.09**	0.18	−0.08	**1.64**
GENERATIVE MODELS															
BEIT-B	0.18	**0.82**	0.02	0.53	**0.65**	−0.09	**−1.02**	0.28	**1.28**	0.09	0.26	**1.14**	**−1.58**	0.56	**1.72**
IGPT-S	0.66	**0.84**	**−1.02**	0.75	0.22	0.16	**−0.55**	**−1.32**	0.54	0.28	0.29	**1.31**	**−1.11**	0.89	**1.69**
VIT-MAE-B	0.11	**0.55**	−0.29	−0.35	**0.59**	0.08	**−1.15**	**−1.15**	−0.81	0.34	0.29	**0.96**	**−1.30**	**−1.31**	**1.75**

Table 4: iEAT effect sizes (see Equation 2) for a range of association tests (see Table 1) using different embedding models. The models were trained on ImageNet-21k using self-supervised methods, with the exception of ViT-MoCo which was trained on ImageNet-1k. The effect sizes indicate the magnitude and direction of the bias, and are written in bold if the effect is significant at $p_t = 0.05$. **ViTs trained using different self-supervised objectives can exhibit opposite social biases, despite being trained on the same dataset.**

Opposite Biases despite same Training Data The analysis of the number of significant biases fails to capture their direction. To address this, we contrast the effect sizes in Table 4. To our surprise, we find that ViTs can exhibit opposite social biases, despite being trained on the same dataset, *e.g.* ViTMAE exhibits a tendency to perceive Native Americans as less pleasant than European Americans, while ViT-MoCo [10] exhibits the inverse association. However, we also find that all models reinforce a handful of consistent social biases irrespective of the training objective, *e.g.* all models associate women more with family roles than careers, and perceive Arab-Muslims as less pleasant than other humans. This points to the idea that these social biases are indeed ingrained from the training data. These findings suggest that biases in image models are not just a result of training data, but that the training objective is a significant factor contributing to their emergence, affecting both the magnitude and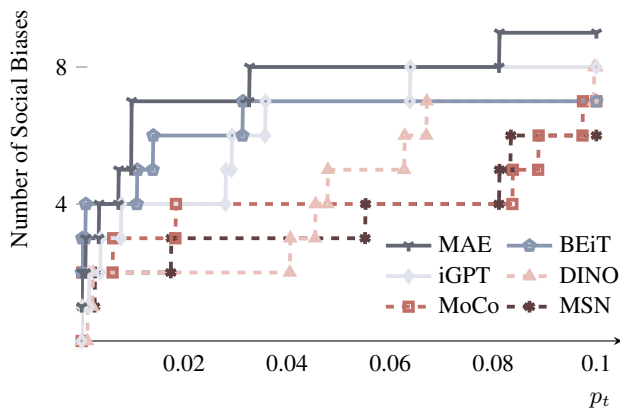 direction of biases. Hence, we suggest that future work on bias mitigation focuses on the set of social biases that is consistent across models.

4.3. Impact of Model Architecture

Model Size The size of a model often impacts its performance, indicating that larger models tend to generate embeddings that contain higher-quality, more general-purpose information about an image. Therefore, we investigate the influence of model scale on social biases, using iGPT [8] and ViT-MAE [13], as both have been trained using self-supervised methods and are available in three different model sizes. The results indicate that as we scale the model scales, the direction of social biases within the embedding spaces remains somewhat consistent (see Table 5). This implies that models of similar architecture, trained on the same dataset using the same training objective, tend to inherit analogous social biases.

Figure 3: The number of biases detected in embedding spaces of ViTs for different values of p_t (see Equation 1). **ViTs trained using discriminative objectives are less biased than those trained using generative objectives.**
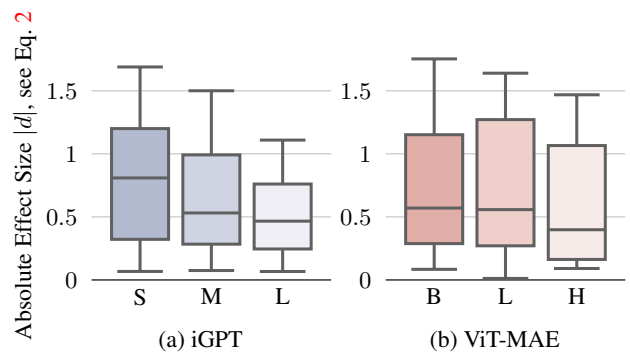
Figure 4: **The mean absolute iEAT effect size decreases as model size increases.** The boxplot illustrates the effect size distribution, with the median (solid line), the quartile range (boxes), and the rest of the distribution (whiskers).

However, we observe that the average magnitude of the social biases decreases as the model size increases (see Figure 4), which implies that scaling the model might be a prac-

MODELS	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
IGPT-S	0.66	**0.84**	**−1.02**	0.75	0.22	0.16	**−0.55**	**−1.32**	0.54	0.28	0.29	**1.31**	**−1.11**	0.89	**1.69**
IGPT-M	0.38	**0.97**	−0.62	0.46	**0.43**	0.19	−0.07	**−1.02**	−0.47	0.60	0.08	**1.26**	0.59	**1.02**	**1.50**
IGPT-L	−0.40	**1.00**	0.41	0.79	**0.44**	0.23	0.27	−0.61	−0.77	0.55	0.07	**1.11**	0.13	0.49	**0.75**
VIT-MAE-B	0.11	**0.55**	−0.29	−0.35	**0.59**	0.08	**−1.15**	**−1.15**	−0.81	0.34	0.29	**0.96**	**−1.30**	**−1.31**	**1.75**
VIT-MAE-L	0.03	**0.56**	−0.21	−0.51	**0.55**	0.01	**−1.17**	**−1.43**	−0.75	0.35	0.33	**1.03**	**−1.38**	**−1.41**	**1.64**
VIT-MAE-H	0.09	**0.63**	−0.39	−0.10	**0.55**	−0.09	**−1.18**	**−1.34**	−0.23	0.29	0.30	**0.95**	**−1.47**	**−1.44**	0.40

Table 5: iEAT effect sizes (see Equation 2) for a range of association tests (see Table 1) using different embedding models trained on ImageNet-21k using self-supervised methods. The effect sizes indicate the magnitude and direction of the bias, and are written in bold if the effect is significant at $p_t = 0.05$. **The direction of the social biases in the embedding spaces of a model are consistent across model sizes. However, the average magnitude of the social biases decreases as model size increases.**

tical strategy to mitigate social biases. We speculate that this could be attributed to the model's capacity to capture more semantic information about the objects in the image, without the need to rely on spurious correlations. However, it is crucial to recognize that scaling a model alone might not be sufficient to eliminate social biases.

Input Resolution and Patch Size In addition, input resolution and patch sizes have been discussed as important model parameters [3, 13]. Hence, we investigate the effect of these parameters on social biases (see Table 6). To assess the impact of different input resolutions, we consider BEiT pre-trained on ImageNet-21k at a 224x224 input resolution and subsequently fine-tuned on ImageNet-1k at different input resolutions. Our results indicate that social biases diminish as input resolution increases. This finding implies that adopting higher input resolution might contribute to a reduction in social biases. To assess the impact of different patch sizes, we consider ViT-DINO [7], which was trained at different patch sizes. In our analysis, we observe some variability in the magnitude of social biases, but

no systematic increase or decrease. However, it's important to acknowledge that the sample size for this analysis is small, due to the limited number of published models. Therefore, further validation should be conducted to confirm these findings.

Per-Layer Analysis In our experiments, we use the embeddings from the layer that has been reported to be optimal in linear evaluation. However, we expect that the intensity of the biases might differ between layers, due to the increasing semantic interpretability of internal representations [5, 33]. To explore this, we determine the number of social biases across different layers, using a significance threshold of $p_t = 0.5$. The results are illustrated in Figure 5. We observe that for models trained using generative objectives, despite some variation in the magnitude, the number of significant biases is somewhat consistent across all layers. However, for models trained using discriminative objectives we find that the number of significant biases in the earlier layers mirrors those of models trained using generative objectives and then decreases as we progress through

MODELS	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
INPUT RESOLUTION															
BEIT$_{224}$-L	**1.59**	**1.41**	0.20	−0.07	**0.40**	−0.21	**1.59**	−0.19	**1.46**	0.18	**−0.88**	**1.12**	**1.06**	0.81	**1.18**
BEIT$_{384}$-L	0.45	**1.46**	0.60	0.15	0.36	−0.17	**1.61**	−0.46	**1.47**	0.27	**−1.11**	0.47	0.61	**1.12**	**1.02**
BEIT$_{512}$-L	0.01	**1.55**	0.35	0.30	0.19	−0.22	**1.65**	−0.41	**1.63**	0.21	**−1.09**	0.49	0.46	**1.03**	**0.79**
PATCH SIZE															
DINO-B/8	0.04	**1.22**	0.32	1.19	**0.37**	−0.16	0.06	**0.97**	**1.16**	0.36	−0.13	0.04	**−1.21**	0.41	**1.49**
DINO-B/16	**0.99**	**1.20**	−0.86	0.88	**0.38**	0.01	−0.12	**0.84**	0.49	0.22	−0.08	−0.13	−0.88	−0.77	**1.24**

Table 6: iEAT effect sizes (see Equation 2) for a range of association tests (see Table 1) of BEiT pre-trained on ImageNet-21k and then fine-tuned on ImageNet-1k at different input resolutions, and ViT-DINO trained using different patch sizes. The effect sizes indicate the magnitude and direction of the bias, and are written in bold if the effect is significant at $p_t = 0.05$. **The direction of the social biases are somewhat consistent across different input resolutions and patch sizes, and the average magnitude of the biases decreases as input resolution increases. However, we do not observe a systematic effect for patch size.**

the model. This suggests that the biases inherent in the low-level features are consistent across all models, but there is a noticeable divergence as the models develop more semantically meaningful features. We hypothesize that the observed divergence in biases across different layers could be attributed to the specific training objectives of the models, as detailed in Section 4.2.

The existence of biases in earlier layers does seem counterintuitive, as no semantic concepts have formed yet. However, we found a substantial portion of these biases, such as skin tone and weight, are connected to lower-level features, such as pixel brightness. This suggests that these biases could be identified without necessarily associating them with the intended semantic concepts. Therefore, we hypothesize that the root of the biases in the earlier layers could be grounded in the inherent characteristics of the image data, and not necessarily the high-level semantic interpretations we are probing. These findings align with prior observations on ResNets [33].
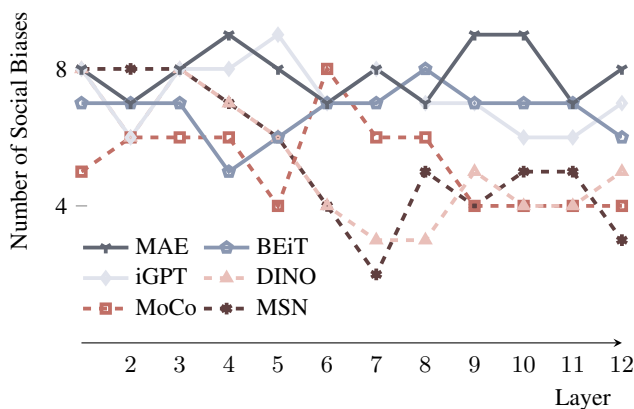

Figure 5: The number of social biases detected across different embedding layers of ViTs using a significant threshold of $p_t = 0.05$ (see Equation 1). **ViTs trained using discriminative and generative objectives share a similar number of biases in earlier layers, but diverge as the models form more semantically meaningful features, such that discriminative models encode less social biases in later layers.**

5. Conclusion

The emergence of social biases in models trained using self-supervised objectives is often attributed to biases in the training data. However, we find that models can exhibit opposite biases despite being trained on the same data. This challenges the prevailing belief that social biases arise just from simple co-occurrences of objects in the training images. Moreover, we find that training objectives, model architecture, and model scale each have significant effects on social biases in learned representations. These effects can

be the reduced, but not eliminated, using counterfactual data augmentation. Therefore, we recommend that model developers and users take these details into account in designing and selecting the model most relevant to their needs, as each decision has quantifiable trade-offs. Moreover, our analysis exposes a set of social biases that is consistent across different models, wherefore we suggest that future work assesses their bias mitigation approaches on these dimensions.

Acknowledgment

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, page 456–473, 2022.

[2] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *International Conference on Learning Representations*, 2020.

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[4] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: Al- locative versus representational harms in machine learning. *9th Annual Conference of the Special Interest Group for Computing, Information and Society*, 2017.

[5] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV 2021 - International Conference on Computer Vision*, pages 1–21, 2021.

[8] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 2020.

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2020.

[10] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629. IEEE Computer Society, 2021.

[11] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[15] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12115–12124, June 2021.

[16] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[17] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Comput. Surv.*, 54(10s), 2022.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR (Poster)*, 2015.

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

[20] Anne Lauscher, Tobias Lueken, and Goran Glavaš. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[22] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022.

[23] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), 2021.

[24] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. Gender artifacts in visual datasets, 2022.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[26] 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.

[27] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 306–316. Association for Computing Machinery, 2021.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[30] 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 (IJCV)*, 115(3):211–252, 2015.

[31] Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, and Novi Quadrianto. Contrastive examples for addressing the tyranny of the majority. *CoRR*, abs/2004.06524, 2020.

[32] Shashank Shekhar, Florian Bordes, Pascal Vincent, and Ari Morcos. Objectives matter: Understanding the impact of

self-supervised objectives on vision transformer representations, 2023.

[33] K. Sirotkin, P. Carballeira, and M. Escudero-Vinolo. A study on the distribution of social biases in self-supervised learning visual models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10432–10441. IEEE Computer Society, 2022.

[34] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 701–713. Association for Computing Machinery, 2021.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[36] 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 (IJCV)*, 2022.

[37] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Adversarial removal of gender from deep image representations. *CoRR*, abs/1811.08489, 2018.

[38] Zeyu Wang, Klint Qinami, Yannis Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. *CoRR*, abs/1911.11834, 2019.

[39] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H. Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. Technical report, 2020.

[40] Ross Wightman. Pytorch image models, 2019.

[41] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 506–523. Springer International Publishing, 2020.

[42] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833. Springer International Publishing, 2014.

[43] George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. Mitigating bias in search results through contextual document reranking and neutrality regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2532–2538. Association for Computing Machinery, 2022.

[44] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning, 2018.

[45] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, 2018. Association for Computational Linguistics.