Shape-Biased Learning by Thinking Inside the Box

Niklas Müller*
Psychology Research Institute
University of Amsterdam
Amsterdam, The Netherlands
n.muller@uva.nl

Cees G. M. Snoek
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
c.g.m.snoek@uva.nl

Iris I. A. Groen†
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
i.i.a.groen@uva.nl

H. Steven Scholte†
Psychology Research Institute
University of Amsterdam
Amsterdam, The Netherlands
h.s.scholte@uva.nl

Abstract

Deep Neural Networks (DNNs) may surpass human-level performance on vision tasks such as object recognition and detection, but their model behavior still differs from human behavior in important ways. One prominent example of this difference, and the main focus of our paper, is that DNNs trained on ImageNet exhibit an object texture bias, while humans are consistently biased towards shape. DNN shape-bias can be increased by data augmentation, but next to being computationally more expensive, data augmentation is a biologically implausible method of creating texture-invariance. We present an empirical study on texture-shape-bias in DNNs showcasing that high texture-bias correlates with high background-object ratio. In addition, DNNs trained on tight object bounding boxes of ImageNet images are substantially more biased towards shape than models trained on the full images. Using a custom dataset of high-resolution, object annotated scene images, we show that (I) shape-bias systematically varies with training on bounding boxes, (II) removal of global object shape as a result of commonly applied cropping during training increases texture bias, (III) shape-bias is negatively correlated with test accuracy on ImageNet while being positively correlated on cue-conflict images created using bounding boxes, following the trend of humans. Overall, we show that an improved supervision signal that better reflects the visual features that truly belong to the to-be-classified object increases the shape-bias of deep neural networks. Our results also imply that simultaneous human alignment on both classification accuracy and strategy can not be achieved on default ImageNet images, suggesting the need for new assessments of both shape-bias and behavioural alignment between DNNs and humans.

1 Introduction

Convolutional neural networks have been successful in learning and performing vision tasks such as object recognition and detection, achieving human-level accuracy (Krizhevsky et al., 2012; He et al., 2016). While humans are known to process and code information efficiently, DNNs have been found to use shortcut learning for task completion (Geirhos et al., 2020; Hermann et al., 2023). Shortcut

*Corresponding author
†Shared senior author

Preprint. Under review.

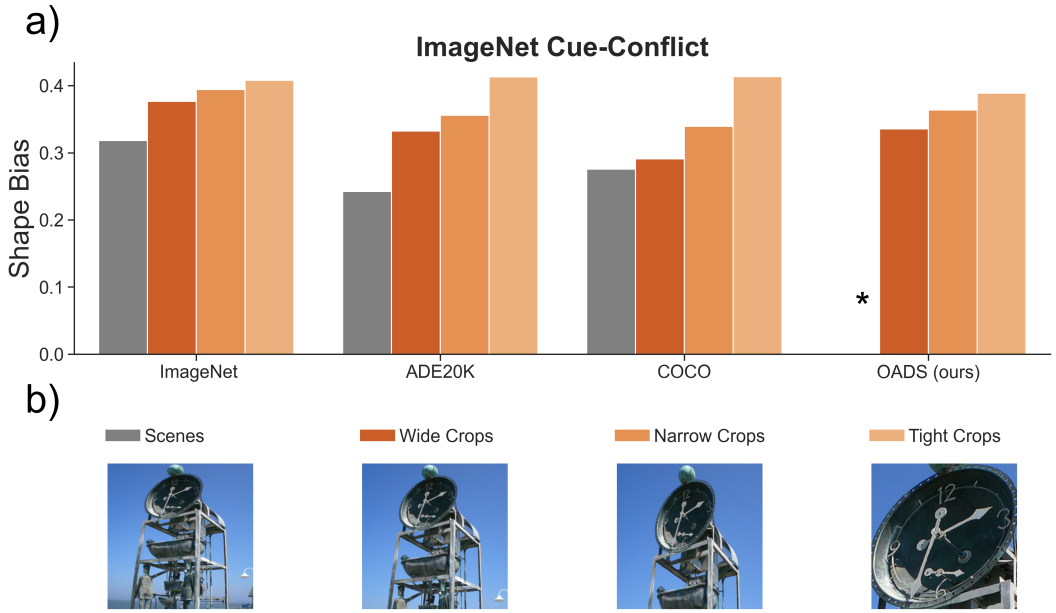

Figure 1: **Tight object crop training increases ImageNet shape-bias.** a) Shape-bias measured using ImageNet cue-conflict images for models trained on different datasets. Models trained on ImageNet are either trained on full images (gray), or on crops created using bounding box annotations available for a subset of ImageNet images using varying degrees of background included (orange shades). Models trained on ADE20K and MS-COCO scenes (gray) are trained on object segmentation and object detection, respectively and then finetuned on default ImageNet images. Models trained on ADE20K, MS-COCO, and OADS crops (orange) are trained on object classification and then finetuned on ImageNet. ($*$ there is no model equivalent to a scenes trained model for OADS). b) Examples for default/scene-like images, wide crops (150% additional background area), narrow crops (80% additional area), and tight crops (no additional area). Example image is from the MS-COCO dataset. Thus, consistent across multiple datasets, shape-bias increases when reducing amount of background included in training crops.

learning allows DNNs to use visual features unrelated to the supervision label that hold significant statistical information across image instances of the same class. Shortcut learning expresses itself through e.g., usage of features that belong to the background instead of the object during object classification training. Further, ImageNet (Russakovsky et al., 2015) trained DNNs have been shown to be strongly biased towards texture when classifying objects while failing to generalized towards object shapes (Geirhos et al., 2018). This texture-bias has been widely discussed as a behavioral discrepancy measure when comparing networks to humans as humans are labelling objects according to their shape (Landau et al., 1988; Ritter et al., 2017). Texture-bias in humans and DNNs is commonly measured with images that contain the shape of one object and the texture of another object, thus exhibiting conflicting cues for classification, here referred to as "cue-conflict images". The further alignment of DNN coding strategies with those of humans holds the promise of more efficient and robust model behavior (Sucholutsky et al., 2023; Jaini et al., 2023; Jang and Tong, 2024; Muttenthaler et al., 2024).

In this paper, we present an empirical study investigating how the background-object-ratio of popular object classification and detection datasets affect DNN texture-bias. We show that training DNNs on tight image crops using bounding box annotations, thereby minimizing the image area occupied by background information, yields increased shape-bias on ImageNet cue-conflict images (see Fig. 1). By reducing the background area, we increase the object-background ratio and thereby improve the supervision signal for object classification training. However, further removal of information, thus likely cropping the global object shape, similarly decreases shape-bias.

As previously observed (Hermann et al., 2020), an increase of shape-bias on ImageNet is often linked to a decrease of classification performance on clean images. There is thus an inherent tradeoff

between shape-bias and classification performance that prevents alignment between humans and DNNs. We constructed a new cue-conflict dataset that is based on high-resolution, tightly cropped object images that resolves the tradeoff between shape-bias and classification accuracy. We thereby demonstrate that humans and DNNs can thus be behaviorally aligned on classification performance and preference for shape over texture.

Concretely, our contributions are: *i)* shape-bias is dependent on object-background ratio used during training, *ii)* classification and generalization performance is negatively correlated with shape-bias on ImageNet, *iii)* humans and DNNs are behaviorally aligned when tested on high-resolution, tightly cropped images.

Overall, we present empirical findings of increased shape-bias that rely solely on the reduction of background-object-ratio and focusing training images on object-related information thereby improving the object classification supervision signal.

2 Related Work

A number of crucial differences in the efficiency of information coding and behavior generation have been pointed out that separate DNNs from humans, including susceptibility to adversarial and random noise or to geometric transformations (Szegedy et al., 2013; Nguyen et al., 2015; Fawzi and Frossard, 2015; Dodge and Karam, 2017) as well as reliance on texture features over object shape. Using images that exhibit conflicting cues for object shape and texture reveals a strong bias of ImageNet-trained DNNs towards texture (Geirhos et al., 2018) and confirms that humans are strongly shape-biased, as previously shown (Landau et al., 1988).

Other work has shown that models with a different architecture, including convolutional models with larger convolutional kernels (Ding et al., 2022) or vision transformers (Tuli et al., 2021) are less texture-biased than other models but are still far from biased towards shape. Conversely, keeping model architecture fixed but instead augmenting training images has been shown to shift model bias towards shape (Geirhos et al., 2018; Hermann et al., 2020). By making texture-features less reliable across repetitions of the same image as well as across instances of the same object class, augmentations can be used to directly reduce the bias towards texture. Augmenting images to express object shape more vigorously has also been found to increase shape-bias (Lee et al., 2022; Gowda et al., 2022). Each of these approaches either manipulates the model architecture, implying that texture-bias is inherent to certain DNN architectures (like the commonly used ResNet50; He et al. (2016)), or manipulates the input data by using augmentations that specifically target certain visual features. Both approaches are considered to be biologically implausible and to not resemble how shape-bias arises in human visual processing. Instead, we show that the object classification supervision signal can be improved by limiting DNN inputs to the smallest bounding box surrounding the object of interest.

Other work has explored DNN texture-bias using representational similarity on ImageNet cue-conflict stimuli instead of object class prediction (Tartaglini et al., 2022). They find that DNNs are only biased towards texture when the transferred texture covers the background and are increasingly biased towards shape when the imposed texture only covers the object but not the background. We share the intuition that high texture-bias in DNNs is closely related to shortcut learning and thus heavy reliance on visual features that do not belong to the to-be-classified object itself. We build upon this work and investigate how not only shape-bias, but also accuracy on cue-conflict images depends on the extent of the imposed texture as well as on the object-background-ratio of training images.

3 Empirical Study

For a set of three commonly used image datasets - ImageNet (Russakovsky et al., 2015), MS-COCO (Lin et al., 2014), and ADE20K (Zhou et al., 2017) - we use deep neural networks trained on several task and assess their texture-shape-bias on ImageNet cue-conflict images (Geirhos et al., 2018) after finetuning each model on ImageNet.

Image Datasets First, we use bounding box information available for a subset of ImageNet images, to create object crops that are more focused on the object and remove object unrelated background information. We then vary to amount of area in the image surrounding the object bounding box that

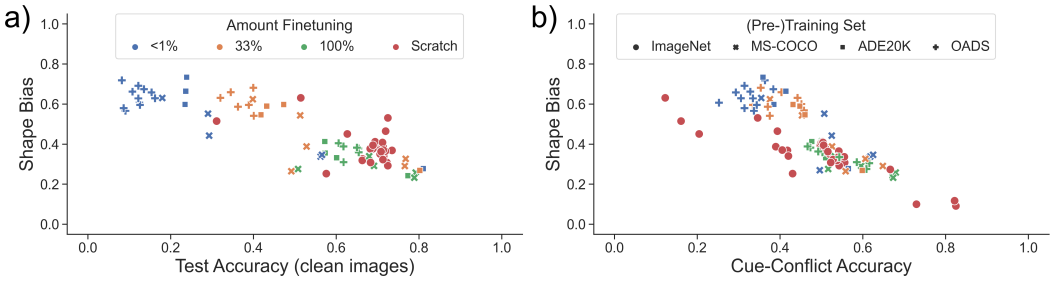

Figure 2: **Shape-bias decreases with accuracy on clean and cue-conflict images.** Test accuracy (clean ImageNet test set; **a**)) vs. shape bias and cue-conflict accuracy (**b**)) vs. shape bias for models pretrained on various datasets and with varying amounts of parameters finetuned on ImageNet. Shape-bias decreases with both test accuracy on clean images as well as generalization accuracy on cue-conflict images (used for measuring shape-bias). The desired model would have both, high accuracy and high shape-bias.

is included in the object crop, from 0% to 80% and 150% of additional surrounding image area that is included (see Fig. 1b).

Second, following the above strategy of creating tight object crops, we use publicly available bounding box annotation for MS-COCO images to create three MS-COCO object crop datasets with varying degrees of bounding box tightness.

Third, we create object crop datasets for ADE20K images using publicly available object segmentation annotations that we use to create object bounding boxes with the minimal possible rectangular area that subtends the fully segmented object. We then create three ADE20K object crop datasets with varying degrees of bounding box tightness.

Last, we use a custom, high resolution scene dataset - Open Amsterdam Data Set (**OADS**) -, consisting of 6130 uncompressed images with a resolution of 5496x3672 pixels. For each scene image, objects have been annotated with a bounding box and one of 16 object class labels (see table 1) resulting in 64.683 annotated objects. We use these bounding boxes to create object crop datasets, again using three degrees of crop tightness.

Training Deep Learning Models For the four object crop datasets (ImageNet, MS-COCO, ADE20K, OADS) described above, we train ResNet50 instances on object classification for 30 epochs. Number of training and validation instances, number of classes, as well as validation accuracy is stated in table 3 for each dataset, respectively.

Additionally, we include models trained on object detection task on MS-COCO (Long et al., 2015), on a scene-segmentation task on ADE20K (Cheng et al., 2021), and models trained on object classification on standard ImageNet images. For the object detection and scene segmentation models, we only use their ResNet50 backbone and replace their task-head with a 2-layer fully-connected object-classification head. We note that for OADS there is no model trained on full scene images (also indicated by the asterisk in Fig. 1a).

We then finetune each pretrained model on ImageNet for 15 epochs. We vary the amount of parameters that are frozen during finetuning and either only train the classification head (<1% parameters finetuned), the last residual block (33%), or the full model (100%).

All custom models were trained using an adam optimizer (learning rate: 0.001) on two NVIDIA A100 and 18 cores, with randomly chosen train and test splits (reported in table 2). A typical training run for object classification on ADE20K was 13 hours, for MS-COCO 3 hours, and for OADS 3 hours. A typical finetuning run on ImageNet took 2 hours.

Texture-shape-bias Assessment Texture-shape-bias assessment for ImageNet is performed on cue-conflict stimuli taken from (Geirhos et al., 2018). Briefly, images consist of conflicting cues for object shape and object texture combined using neural style transfer (Gatys et al., 2016). Essentially, images show the shape features of one object and the texture features of another. Classifying these images can thus reveal a preference for texture or shape.

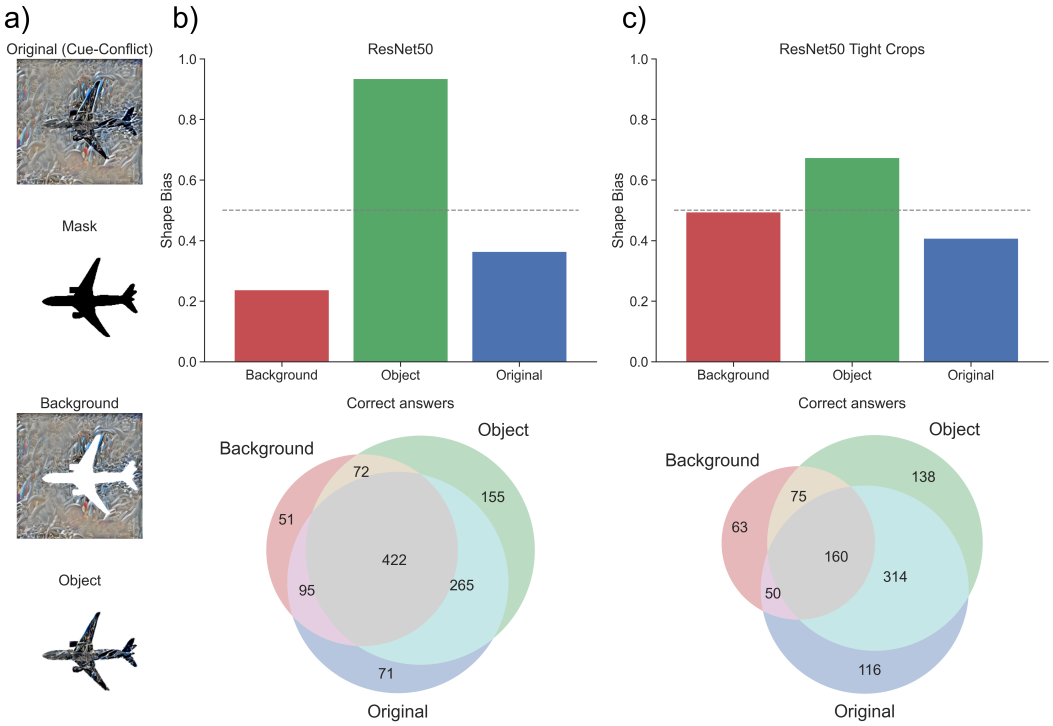

Figure 3: a) Example of original cue-conflict images, object silhouette used to mask the object, texturized-background with masked object, and texturized-object with masked background (from top to bottom). b) Shape bias for the default ImageNet-trained ResNet50 when using the original cue-conflict images, or the object or background masked images (top). Venn diagram showing the number and overlap of correct answers for each masking condition (bottom). c) Same as b) but for a model trained on tight ImageNet object crops using bounding box annotations. Thus, even though shape-bias increases when tested on masked objects, the model is still classifying images according to their texture, when tested on masked backgrounds. This texture usage is reduced when trained on tight object crops instead.

To test texture-shape-bias independently from ImageNet classification ability, we created a new set of cue-conflict images for OADS using the same methodology and hyperparameters as for ImageNet cue-conflict stimuli from (Geirhos et al., 2018). However, *OADS cue-conflict* images are created with a resolution of 400x400. From a set of OADS images for which the standard OADS-trained ResNet50 had 100% classification accuracy, we randomly drew 62 images and performed neural style transfer for each pair of images. We randomly selected 880 from the resulting 3844 images, while maintaining the balance between number of images per class. We created 9 aggregate classes (see italic classes in table 1), e.g., merging the classes "Front door" and "Balcony door" into one "Door" class which are subsequently used for classification.

Cue-Conflict Image Masking Using silhouette images from Geirhos et al. (2018) and the methodology outlined in Tartaglini et al. (2022), we test shape-bias when masking cue-conflict images to either only show the object or only the background, while masking the respective other area. The imposed cue-conflict is thereby maintained, however limited to the extent of the object or the background. We evaluate both accuracy and shape-bias for the original cue-conflict images, the object-masked cue-conflict version, and the background-masked cue-conflict version (see Fig. 3a).

Human Behavioral Experiment We performed a human experiment using the OADS cue-conflict images to assess human shape-bias. We replicated the experimental paradigm from (Geirhos et al., 2017) as much as possible. Specifically, 14 participants each performed 880 trials, where each trial consisted of the presentation of a central fixation cross for 300 ms, followed by the presentation of a randomly drawn, non-repeating OADS cue-conflict image for 200 ms, followed by a pink noise

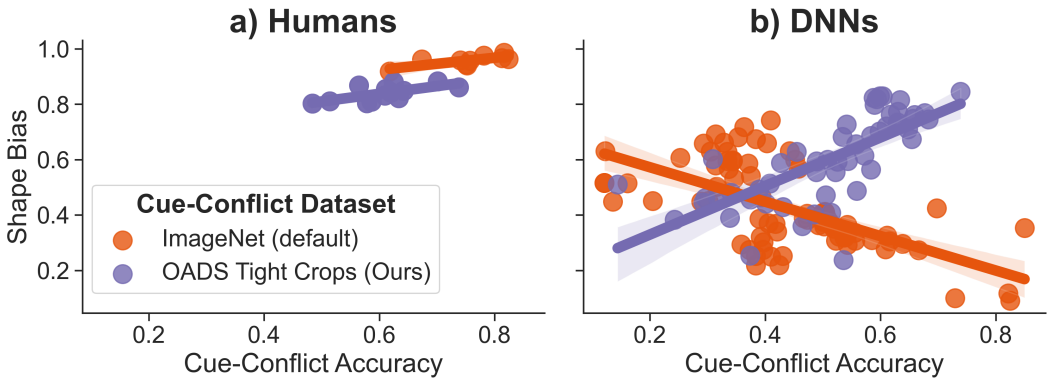

Figure 4: **Texture bias assessment is human aligned when using tightly cropped cue-conflict images.** Accuracy (on cue-conflict images) vs. shape-bias of humans **a)** and DNNs **b)** for ImageNet (orange) and OADS (ours; purple). DNN shape-bias is negatively correlated with cue-conflict accuracy on ImageNet but positively on OADS as well as for humans on both cue-conflict datasets.

mask for 200 ms, followed by a 2500 ms time window to give a response in the form of a mouse click on one of 9 object class names presented on the screen. Participants were introduced to the set of classes before the start of the experiment, and order and arrangement of classes did not change throughout the experiment. Further, performance of participants (calculated as accuracy, counting both texture and shape class as correct) was presented to the participant every 5 trials. After every 50 trials, participants were given a break for at least 30 seconds and self-initiated the next set of trials. Participants were rewarded with research credits or an equivalent monetary reward. Collection and use of human data was approved by the institution's IRB (more details in A.1).

Shape-bias was calculated as the number of shape decision over the sum of shape and texture decisions, for both, DNNs and humans. DNN decisions are based on the class associated with the maximally activated output node. Additionally, cue-conflict accuracy is calculated as the sum of shape and texture decisions over the total amount of images, and is throughout this study referred to as generalization performance. We excluded one participant that had an accuracy of more than two standard deviations away from the population mean.

4 Results

Training on Tight Object Crops Increases DNN Shape-Bias We use multiple versions of ImageNet that feature varying degrees of bounding box tightness, ranging from maximally tight object crops (0% surroundings), over narrow crops (80%) to wide crops (150%). We show that shape-bias on ImageNet significantly increases when training on tight object crops using bounding box annotations. Shape-bias then decreases in a step-wise fashion with decreasing bounding box tightness (see Fig. 1).

High-Quality Object Annotations Can Yield Human-Level Shape-Bias Pre-training models on image datasets that feature object bounding box annotations on larger (scene-)images allows for controlling the amount of background that is included during object classification training. We use two common image datasets, MS-COCO and ADE20K, as well as a custom, high-resolution scene dataset - **OADS** - and train multiple model instances using varying degrees of crop tightness (test accuracy listed in table 3). We compare these models to publicly available models trained on an object detection task for MS-COCO and object/scene segmentation for ADE20K. We observe a strong trend of increasing shape-bias when training on more tightly cropped object images. With additional surrounding area included in the object crop during training, the model's final shape-bias decreases. For models pre-trained on object detection and segmentation on MS-COCO or ADE20K, we find no increase of shape-bias after finetuning on ImageNet compared to default ImageNet-trained models (see gray bars in Fig. 1a).

Next, we want to assess whether models trained on MS-COCO, ADE20K or OADS are inherently more shape-biased and become texture-biased through finetuning on ImageNet. Instead of finetuning the full model, we keep a certain amount of parameters fixed during finetuning, thereby maintaining

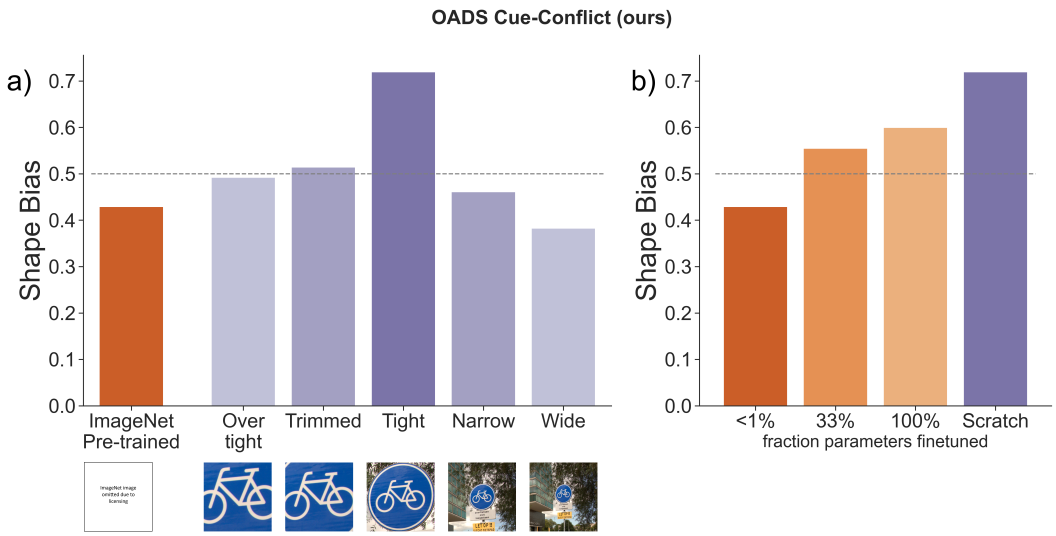

Figure 5: **Controlling DNN shape bias** a) Shape bias for models pretrained on ImageNet and finetuned on OADS (orange) and models trained from scratch on OADS (purple) using varying degrees of object-background-ratio. b) shape bias for models with varying amounts of parameters finetuned on OADS (orange) vs. a model trained from scratch (purple).

the convolutional features from the pre-training. Fig. 2a) shows that finetuning only the classifier layer does not yield a well performing model on clean images but indeed yields models that are strongly shape-biased. With increasing amount of parameters that are finetuned on ImageNet, clean image test accuracy gradually increases while shape-bias decreases. Fully finetuning models on ImageNet leads to the same level of texture-bias as training from scratch.

The negative correlation between accuracy and shape-bias is even more pronounced when evaluating accuracy on the cue-conflict images (see Fig. 2b). In order to show high generalization accuracy on cue-conflict images, models need to be biased towards texture.

Finegrained Texture-Bias Assessment Training models on tightly cropped objects increases shape-bias measured by ImageNet cue-conflict images. A critical features of these images is that the transferred texture is spread out across the full image. Thus, models that classify objects by using background features might be more texture-biased using this assessment strategy. We ask whether the increased shape-bias of models trained on tight object crops also reduces the reliance on background features.

ResNet50 trained on default ImageNet images shows a strong shape-bias when the imposed texture is limited to the extent of the object silhouette (see Fig. 3b). In contrast, when masking the object and keeping the texturized background, models are strongly texture-biased. We replicate this finding from Tartaglini et al. (2022). Crucially, when comparing the amount of correct answers, thus always counting both the texture and the shape label as a correct prediction, models perform equally well on the original, the masked-object, and the mask-background cue-conflict images. Thus, the default ImageNet-trained model classifies cue-conflict images equally well using shape or texture cues.

In contrast, the model trained on tight object crops shows an increase shape-bias on the original cue-conflict images, as shown in Fig. 1. Interestingly, we see that the discrepancy of shape-bias between masking the object or the background is reduced (see Fig. 3c). Additionally, this model performs significantly worse when only presented with the texturized background as compared to being presented with the texturized object, thus showing a stronger tuning to the cue within the extent of the object instead of the background.

Resolving the Accuracy-Shape-Bias Tradeoff Using our custom, high resolution image dataset **OADS**, featuring manually annotated bounding boxes, we created a cue-conflict image dataset using the same methodology, but that is instead based on tight object crops: *OADS cue-conflict*.

First, we performed a human experiment and replicate the strong shape-bias of humans on our custom cue-conflict dataset (see Fig. 4a). Second, we assess the relationship between cue-conflict accuracy and shape-bias for humans using the data from Geirhos et al. (2017) and our own experiment. For both cases, we observe a positive correlation between cue-conflict accuracy and shape-bias. Third, as outlined in Fig. 2b), models that perform well on classifying cue-conflict images due so using texture, while shape-bias increases when accuracy decreases. Thus, this poses an additional misalignment between human behaviour (positive correlation) and ImageNet-trained DNN behavior (negative correlation). Fourth, assessing OADS trained or finetuned models on OADS cue-conflict images reveals a positive correlation between cue-conflict accuracy and shape-bias, thus strongly aligning with humans.

Controlling DNN Shape-Bias DNN shape-bias approaches human-levels when both trained and assessed on tightly cropped objects. We now ask whether increasing the amount of background included in object crops during training also affects DNN shape-bias when tested on tightly cropped cue-conflict images. Fig. 5a) shows that inclusion of background leads to a stepwise decrease of shape-bias (see light purple bars in 5a). Crucially, these models have an equally low shape-bias as an ImageNet-trained model that is finetuned on OADS (see orange bar in Fig. 5a).

Further, removing all background information by tightening the bounding box even more, thus also cropping away parts of the object itself leads to models that are neither texture nor shape biased.

As hypothesized before and shown in Fig. 2, we expect models trained on tight object crops to be shape-biased that then become texture-biased through finetuning on ImageNet. Using the default, texture-biased ImageNet-trained model, we apply the same methodology of freeze parts/most of the network's parameters while finetuning the other parameters on OADS (see Fig. 5b). Finetuning only the classifier layer (thus <1% of all parameters) yields a shape bias of 42% while already achieving a test accuracy on clean images of 80%. Increasing the amount of finetuned parameters to 33% and 100% yields a stepwise increase of shape-bias. However, even the fully finetuned model does not reach the same shape bias as the model trained from scratch.

To summarize, we find that OADS offers important qualities for further alignment of DNN and human behavior. Crucially, this yields models with more efficient information coding strategies that emerge from the use of more ecologically valid training data and rely less on auxiliary methods, e.g., image cropping or data augmentation.

5 Discussion

In this paper, we present an empirical study examining the influence of background-object-ratio plays on DNN texture-shape-bias. We systematically manipulate the amount of background visible in each object image in three commonly used image datasets (ImageNet, MS-COCO, ADE20K) and assess the effect on texture-shape-bias as evaluated using cue-conflict images.

Images that present ecologically valid scenes, i.e. objects that occur together in the real visual world in a scene/on a background that holds clues about the object's identity increases the likelihood of shortcut-learning in DNNs. Features of the background that are shared across many instances of the same object class can be utilized for object classification. However, this might result in low generalization abilities for cases where the same object occurs in a non-ecologically valid background or scene. Further, the focus on background may also increases the sensitivity to high spatial frequencies and therefore to textures. Here, we show that reducing images to the minimum rectangular area that is actually occupied by the to-be-classified object can thus lead to less shortcut learning and bias DNNs towards shapes instead of textures.

There has been a lot of interest in the discrepancy of texture-/shape-bias between humans and DNNs both from the perspective of computer science and representation learning (Geirhos et al., 2021; Hermann et al., 2020; Li et al., 2024; Orhan and Lake, 2024) as well as in human neuroscience (Kim et al., 2019; Jagadeesh and Livingstone; Subramanian et al., 2024). So far, the most promising work showing a decrease in texture-bias used strong image augmentations (Hermann et al., 2020; Jang and Tong, 2024). These augmentations often target image features that are related to texture (color jitter, high-spatial frequencies, etc.) thereby making these features less useful for object classification training. However, this merely results in a decrease in texture-bias but does not necessarily bias the model towards shape, as models are still capable of classifying images by their texture. In terms

of human alignment, a decrease in texture-bias through augmentation is not biologically plausible. Instead, through 3D vision, motor-interaction and supervision by adults, infants learn about the strict boundaries of objects, do not necessarily rely on the background information accompanying the object, and can easily recognize objects in isolation. Removal of object unrelated background information is thus a way of improving the supervision signal for DNNs during training. The model is thus left with information that is to the majority bound to the actual object for which it needs to learn a label.

Generally, our results show that if only object-bound information is present, DNNs develop a shape-bias, as measured by cue-conflict images. Once more information that is not object-bound (background) is part of the training images, DNNs start relying more on texture features for classification. However, further cropping - often applied during DNN training - can remove parts of the global object shape which we show leads to a decrease in shape-bias. This emphasizes again that DNNs rely on available over relevant visual features (Hermann et al., 2023).

Interestingly, we also show that models that are trained on scene segmentation and thus also segment objects from their background, do not develop a shape-bias. However, models trained on object classification using object crop from the same dataset are increasingly more texture-biased when training images contain more background. There is thus a discrepancy in whether a model is capable of defining the extent of an object (during segmentation) or whether an association between visual information and a label needs to be learned.

Overall, we see a strong negative correlation between model's performances (both on clean and cue-conflict images) and shape-bias when evaluating models on ImageNet. This poses a crucial problem for the study of human visual processing using deep neural network models: behavioral alignment can only be achieved for classification performance or for classification strategy, but not both. However, we find that when shifting both the training image distribution as well as the test images to include less background information and be more focused on the objects, humans and models achieve similar performances using the same strategy, thereby paving the way for more human-like models without a need for altering training regimes (e.g. stylized training Geirhos et al. (2021)) or augmentation Hermann et al. (2020); Jang and Tong (2024).

Altogether, our findings support previous findings that high texture-bias is not a necessary property of deep learning models but is tightly linked to the training data and therefore the way they see. This differs significantly from human visual perception thus likely leading to different bias in both behavior and internal representations. We show that simple manipulations, using existing bounding box annotations, can help shift the training data distribution to be more aligned with human vision.

Acknowledgments and Disclosure of Funding

This work was funded by the University of Amsterdam Data Science Center Interdisciplinary PhD Program.

References

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022.

Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pages 1–7. IEEE, 2017.

Alhussein Fawzi and Pascal Frossard. Manitest: Are classifiers really invariant? *arXiv preprint arXiv:1507.06535*, 2015.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

Robert Geirhos, David HJ Janssen, Heiko H Schütt, Jonas Rauber, Matthias Bethge, and Felix A Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*, 2017.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

Shruthi Gowda, Bahram Zonooz, and Elahe Arani. Inbiased: Inductive bias distillation to improve generalization and robustness through shape-awareness. *arXiv preprint arXiv:2206.05846*, 2022.

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*, pages 770–778, 2016.

Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020.

Katherine L Hermann, Hossein Mobahi, Thomas Fel, and Michael C Mozer. On the foundations of shortcut learning. *arXiv preprint arXiv:2310.16228*, 2023.

Akshay Vivek Jagadeesh and Margaret Livingstone. Texture bias in primate ventral visual cortex. In *ICLR 2024 Workshop on Representational Alignment*.

Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779*, 2023.

Hojin Jang and Frank Tong. Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. *Nature Communications*, 15(1):1989, 2024.

Taekjun Kim, Wyeth Bair, and Anitha Pasupathy. Neural coding for shape and texture in macaque area v4. *Journal of Neuroscience*, 39(24):4760–4774, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988.

Sangjun Lee, Inwoo Hwang, Gi-Cheon Kang, and Byoung-Tak Zhang. Improving robustness to texture bias via shape-focused augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4323–4331, 2022.

Tianqin Li, Ziqi Wen, Yangfan Li, and Tai Sing Lee. Emergence of shape bias in convolutional neural networks through activation sparsity. *Advances in Neural Information Processing Systems*, 36, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. *Advances in Neural Information Processing Systems*, 36, 2024.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

A Emin Orhan and Brenden M Lake. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, pages 1–13, 2024.

Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning*, pages 2940–2949. PMLR, 2017.

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.

Ajay Subramanian, Elena Sizikova, Najib Majaj, and Denis Pelli. Spatial-frequency channels, shape bias, and adversarial robustness. *Advances in Neural Information Processing Systems*, 36, 2024.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Alexa R Tartaglini, Wai Keen Vong, and Brenden M Lake. A developmentally-inspired examination of shape versus texture bias in machines. *arXiv preprint arXiv:2202.08340*, 2022.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

A Appendix

A.1 Instruction for participants

Human participants were instructed to perform the experiment as follows: *In this experiment, you will perform a visual decision-making task. You will see an image and will have to answer what kind of object you have seen in the image. You give your answer by clicking on one of the 9 object names on the screen. In each trial, an image will be presented for a short amount of time, then you will see a noise pattern, and after that you will see the overview of object names where you can choose one to click on. The next trial will follow immediately. Every 50 trials there will be a break of at least 30 seconds, after which you can decide when to continue with a mouse click. We will track your performance and display your performance score every 5 trials. Please try to score as high as possible. Note that for every image there are multiple possible answers. The experiment consists of 880 trials and will last about 75 minutes, depending on the length of the breaks that you choose.*

Participants were compensated monetarily with 12.5 in the local currency or an equivalent of 1.5 research credits.

A.2 Dataset and Training Information

OADS Object Class Labels			
Traffic light	*Tree*	*Car*	*Bench*
Van	Balcony door	*Scooter*	*Truck*
Bollard	Traffic sign	*Bike*	*Bin*
SUV	Carrier bike	*Front door*	*Lamppost*

Table 1: OADS object classes

	MS-COCO	ADE20K	OADS
Train set	688000	507476	57496
Validation set	86001	73129	7187
Number of classes	80	3542	16

Table 2: Number of instance in training and test set and number of classes

	MS-COCO	ADE20K	OADS
Tight Crops	0.7528	0.5508	0.9186
Narrow Crops	0.7297	0.5439	0.9144
Wide Crops	0.6816	0.4706	0.8902

Table 3: Top-1 test accuracy of ResNet50 after training of 30 epochs