Exploring Shape and Texture Bias in Object Detection

Yejin Kim, Junha Kim, Doyeon Lee
Graduate School of Data Science, Seoul National University
{a2000yejin, terajunha, doyeon.lee}@snu.ac.kr

Abstract

Ways to mitigate CNN's texture bias has been studied in various works under the image classification task. However object detection models were not explicitly examined whether texture bias is really an issue. We hypothesized that object localizing task and its corresponding loss helps the model learn more shape related features since the bounding box needs to be aligned with the silhouette of the object. In the preliminary experiment we found previous augmentation methods used to test texture bias of classification models were not quite adequate in measuring texture/shape bias. Thus we modified several augmentation methods to suit the object detection task. We tested three current State-Of-The-Art object detection models on augmented images and found that texture bias was low in all three models. We suggest injecting the idea of localization to image classification models to increase robustness against texture changes. Also, further work on other detection models using various augmentation methods will be needed to get a better understanding of how models work and develop features that are not susceptible to texture deformed images, such as paintings.

1. Introduction

Object detection is a major computer vision task along with image classification and semantic segmentation. Object detection really started to take off with the introduction of Convolutional Neural Networks(CNN) developed by AlexNet [17] under the image classification task. State-of-the-art object detection models such as Faster R-CNN [27] and YOLO [26] mainly follow standard CNN architecture (e.g. AlexNet, VGG) and DETR [3] employs a transformer with CNN extracted features.

CNN was thought to use high-level features such as shapes to understand images but recent works suggest the opposite. [11] claimed that unlike humans, CNN relies on simple correlations such as texture rather than shapes to classify images, making it susceptible to perturbations and domain shifts. Other works [1, 32] also support CNN's bias towards texture.

While various methods have been proposed to reduce CNN's texture bias by enhancing shape bias [11, 18], some argue that shape bias is mostly unrelated with corruption robustness [22, 25]. However, each work uses different datasets making it hard to verify their arguments. [15] also felt the need for a common dataset and created six datasets for measuring texture and shape bias but our work differs in that it is designed for object detection instead of image classification. CNN model's texture bias has been studied mainly under the image classification task but we believe a separate study is needed for object detection since it is trained with more shape information to correctly localize objects.

Also, each existing dataset's ability to properly capture the object feature (e.g. shape, texture, color) it was supposed to is questionable considering that some augmented images have background information while some don't. We believe the background can be a confounder and created a total 13 datasets of 5 augmentation methods with three variations – no background ('mask'), original background ('obj'), and augmented background ('all') (edge and silhouette methods only have two variations). We found that there is a significant difference in performance between images with different background variation. However, we discovered several limitations of existing augmentation methods especially when applied to object detection. Therefore, we chose three additional augmenting methods resulting in 11 datasets of 4 styles with three background variations each (except for patch shuffling which has two variations).

By testing SOTA object detection models on these datasets and qualitatively examining the results using saliency maps, we aim to understand the effect of shape and texture bias in object detection models. Our contributions are as follows.

- We propose 11 datasets of 4 styles - style transfer, cue-conflict, color compression, and patch shuffle - as a common dataset to measure texture bias in object detection models.
- We found that current object detection models are more sensitive to shape changes than texture changes. We be-

lieve the supervision provided by bounding box labels enhance shape bias.
- Based on our finding, we propose two new ways to enhance shape bias in the classification task. First is through creating an dataset that contains one object per image and the second is through an additional loss term that forces the model to look at similar shapes in texture deformed images.

In section 2 we review previous works which try to alleviate texture bias in image classification and object detection models through data augmentation. In section 3 we report the problems found in existing augmentation methods. Based on this preliminary finding we create new datasets and interpret the results of three SOTA models in section 4. Lastly we summarize our study and suggest new ways for future research.

2. Related Work

2.1. CNN's texture bias

Although CNN based models have achieved outstanding performance in photo realistic datasets (e.g. ImageNet [28], MS COCO [19], PASCAL VOC [7]), their accuracy on out-of-domain (OOD) data such as noisy images and paintings is significantly lower [14, 26]. Fine-tuning models directly on target domains such as People-Art [2, 31] still fell short of models trained and tested on photographs. [11] claimed that CNN's bias toward texture is the reason behind its susceptibility toward corruption. By fine-tuning a CNN model on images stylized by paintings (Stylized ImageNet), they achieved 81% accuracy in the texture-shape cue conflict dataset where only texture was changed to that of other classes (e.g. cat shape with elephant texture). Following this study, many other works used style transfer or other data augmentation methods (e.g. color jitter, random greyscale, random Gaussian Blur, weather effects, patch shuffling) [12, 14, 16].

Other works added auxiliary tasks such as predicting rotation [12] and style [5]. Architectural changes such as adversarial training [33] and contrastive learning [18] were also used to mitigate CNN's texture bias.

2.2. Data augmentation

Most works focus on data augmentation as a way to measure and lessen CNN's texture bias. With images created using various augmenting techniques they train standard CNN models. The role of each augmenting technique is interpreted by the authors [18] or evaluated using shuffled image patches and texture-shape cue conflict images [13, 22]. They also test corruption robustness of fine-tuned models using stylized images, artworks [13], and images augmented by Gaussian blur, phase noise, weather types, contrast changes etc. [11].

These studies all view data augmentation as a useful tool for debiasing CNN models but whether enhancing shape bias leads to corruption robustness is still debatable. [22] found that models trained on edge images showed high shape bias but low robustness to image perturbations. They argued that shape bias doesn't lead to corruption robustness and that learning a robust representation through style variation whether it be shape related or not is more important. Their claim is contradictory to other works [11, 14, 18]. However, the datasets used for training and testing are different across papers, making it hard to reach a concise conclusion. For example, edge images used in [11] excludes the background but [22] does not. Also, [11] applies style transfer to the entire image including the background which can lead to background bias. We felt the need to create a common dataset. We created 11 datasets of 4 styles with 3 variations each (except for patch shuffling) to test each augmentation method's influence on model performance while also controlling the background.

2.3. Object Detection

Object detection has a lot in common with classification tasks since they both use CNN backbones trained on ImageNet. Thus, object detection models such as Faster R-CNN, YOLO and DETR are also presumed to have the problem of texture bias. Based upon this assumption several works finetune object detection models on augmented datasets. However, testing whether object detection models really do have the problem of texture bias is needed since these models are trained differently from classification models. We hypothesize that object detection models are more shape focused since they require a bounding box to be aligned with the object's outline shape [11].

Based on the assumption that object detection models are biased towards texture several works examine their performance in OOD datasets. [2] and [31] used VOC2007 and People-Art dataset to finetune R-CNN and Fast R-CNN respectively. [14] transferred the styles of paintings from the Painter by Numbers dataset directly to the MS COCO dataset. They trained Faster R-CNN on this StyleCOCO dataset and achieved 0.68 AP50 on the People-Art dataset.

Our work applies various augmentation methods to ImageNet-S which has instance level annotations and MS-COCO. We use these to test three SOTA models without fine-tuning to examine whether texture bias is a real problem even in current detection models. We chose CO-DETR, DiffusionDet and YoloX as SOTA models for the following reasons. CO-DETR is the SOTA model in the COCO test-dev dataset scoring the highest mAP. DiffusionDet employs a diffusion model which is known

to reduce classification error in images with disrupted texture [6] and we wish to see if this still holds under the object detection task. YoloX is an extension of the Yolo series which strictly follows a CNN structure compared to CO-DETR and DiffusionDet and its performance surpasses YoloV.4 and V.5. We detect 16 classes in ImageNet-S and 80 classes in MS-COCO.

3. Preliminary experiment

In the following section, we introduce preliminary experiments designed to explore the differences in shape and texture bias between image classification and object detection using the ImageNet dataset.

First, we present the ImageNet-S dataset used for detection (Section 3.1). We then explain how we applied existing experimental methods to augment this dataset (3.2). Finally, we analyze the results (3.3) and discuss the limitations (3.4) demonstrating the necessity for our newly designed experiments in Section 4.

3.1. Dataset

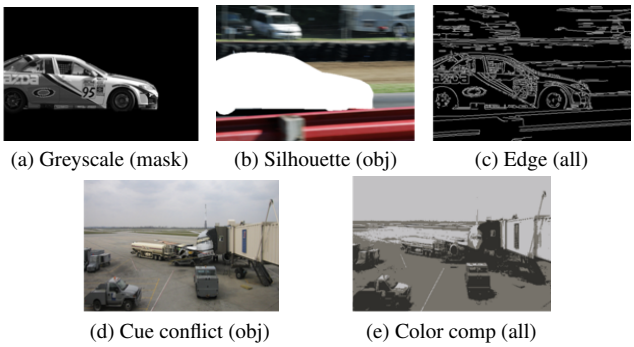
CNN's previous research on texture bias was created with an ImageNet-based dataset [28]. In order to compare the texture/shape bias of detection models with existing classification models, we used ImageNet images grouped to 16 upper classes with WordNet hierarchy [20] same as previous researches [11]. And for the detection task, we made bounding boxes from class segmentation information provided by ImageNet 1000-based ImageNet-S dataset [8].

However additional patch merging was needed. Objects of the same class but separated segments are merged into one box. This preprocessing was only applied to ImageNet-S data which doesn't have ground truth bounding boxes, and was not applied to the MS-COCO dataset.

3.2. Augmentation methods based on image classification

In the preliminary study, we applied 4 augmentation methods used to test texture bias of classification models and 1 additional method we designed. The 4 methods widely used in previous classification works are 'greyscale' which eliminate color and 'silhouette', 'edge' and 'cue-conflict' which distort texture. Specifically, 'cue-conflict' applies the texture of another class onto the original image. The method we designed is 'color-compression' which aims to remove texture while preserving shape.

We believe the background information can help the model detect the object regardless of its augmentation type. Thus we control the background by creating three variations of background deformation - 'all', 'obj' and 'mask' - for each augmentation. 'All' applies deformation to the entire original image, 'obj' transforms only the object while leaving the background unchanged and 'mask' removes the background completely transforming only the object. Sample images for each style with a specific background type can be found in Figure 1.



(a) Greyscale (mask) (b) Silhouette (obj) (c) Edge (all)

(d) Cue conflict (obj) (e) Color comp (all)

Figure 1. Data augmentation

3.3. Result

In the dataset we created, we combined the ground truth labels of overlapping objects into one bounding box. If several small bounding boxes are detected, it is highly likely that they are separate parts of a single object. Therefore, to prevent the detected bounding boxes from being calculated with a small IOU, predicted boxes are also combined into one bounding box. This was applied only when the ratio of the overlapping region between each detected box and the ground truth box to the ground truth box exceeded the specified threshold ratio.

For accuracy assessment, we considered mAP (mean Average Precision), accuracy, and shape-texture ratio representing proportion of correct answers aligned with either shape or texture. Additionally, we analyzed the results through qualitative analysis of bounding box predictions to further understand the outcomes.

Low texture bias in detection models. Consistent with prior research, we measured the ratio of correctly identified answers based on shape versus texture in cue-conflict images. As seen in Figure 2, the average proportion of texture-based correct answers was significantly lower in 'obj' and 'mask' dataset, indicating the absence of texture bias in these datasets. Although there were more texture-based correct answers in the 'all' dataset, the ratio was still lower than several classification models. Considering the data augmentations used here had some limitations in capturing shape and texture bias - more on this in section 3.4 - we used other augmentation methods to validate this claim - more on this is section 4.

Difference between classes. A general decrease in mAP was observed through data augmentation. When classifying this trend based on their accuracy propensity, three different

Figure 2. Ratio of texture-based correct answers per class

The dashed line is the average ratio.

Figure 3. Ambiguity of cue conflict(texture deform) methodology

patterns were seen. This led us to speculate that the ratio of reliance on shape or texture might vary across different classes.

Background influence It appeared that YoloV8 heavily relied on the original background. When the background was removed or augmented, performance significantly deteriorated regardless of the augmentation type. This signifies that the background can be a confounder when measuring texture/shape bias and the effect of augmentation methods should be compared within the same background variation. Further analysis of the influence of background is warranted.

3.4. Limitation

We tested only Yolov8, a common one-stage model which conducts region proposal and classification simultaneously, enabling quick detection with high throughput.

As seen in Figure 3, the previous cue-conflict method was not suitable for accurately measuring individual biases since the so-called 'texture' image still has shape information of small objects. Looking at the predicted bounding boxes and saliency map(CAM method) [21], we concluded that this kind of texture is especially problematic in object detection. Since an image with many repetitions of the same object is not being interpreted as texture by the model, we will narrow down the meaning of texture to full-width patches of an animal or material in the main experiment. More details are provided in Section 4.

Also, we found the quality of edge and silhouette transform was very different depending on which original image was used. Previous works interpreted edge images as containing shape details but we excluded edge and silhouette datasets in our main experiments since they remove too much information in some images.

4. Experiment

Building on the issues identified in existing image augmentation methods for measuring shape and texture bias, we propose a new augmentation method (Section 4.1). We applied this method to the COCO dataset, which is commonly used for object detection. Additionally, we selected the latest state-of-the-art(SOTA) models to investigate whether bias still impacts their performance (4.2).

Our experiments revealed that all tested models exhibited a greater shape bias than texture bias, with detailed results discussed later (4.3). Finally, based on insights gained during the experiments, we suggest several directions for future experiment (4.4).

4.1. Augmentation methods

Style transfer image. Style transfer[9] uses the VGG network [29] to deform original images into art reference image styles. Previous works randomly selected one artwork from the entire Painters by Numbers dataset [30] but we randomly selected from only 8 paintings to have more control on the quality of the augmented images.

When generating images, the VGG network applied gradient descent using the L-BFGS algorithm instead using Adam. This variation has all, obj, and mask images.

Cue conflict image. We used the same VGG network used above to transfer styles of texture reference images to the original image since simply blending texture images like we did in the preliminary study preserved the texture of the original image. Also, in the preliminary study, texture images were created based on the training dataset's class, but because images of repeated small objects had too much shape information, we only used real texture-like images (cat fur, dog fur, elephant skin, rubber tires, wood bark, paint etc.).

For each texture image two non-copyrighted images were selected using Google's image search, and two more texture images were created through prompt engineering using DALLE-3 in ChatGPT-4[23, 24] just like the method used in the preliminary study. Each texture style has four images and is randomly applied to the original image to be transformed. This variation has all, obj, mask images.

Color compression image. Original image was deformed by using the cv2.kmeans function. We set the cluster parameter to 3. We also experimented with 20 clusters in

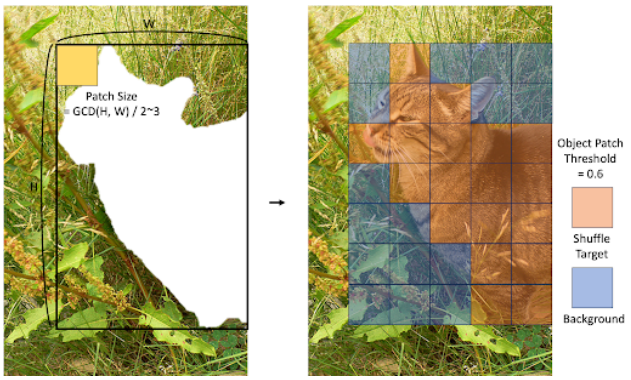


Figure 4. Patch shuffling for object detection

the preliminary study, but this was excluded because there seemed to be no significant difference with the original image. This augmentation dataset has all, obj, mask images.

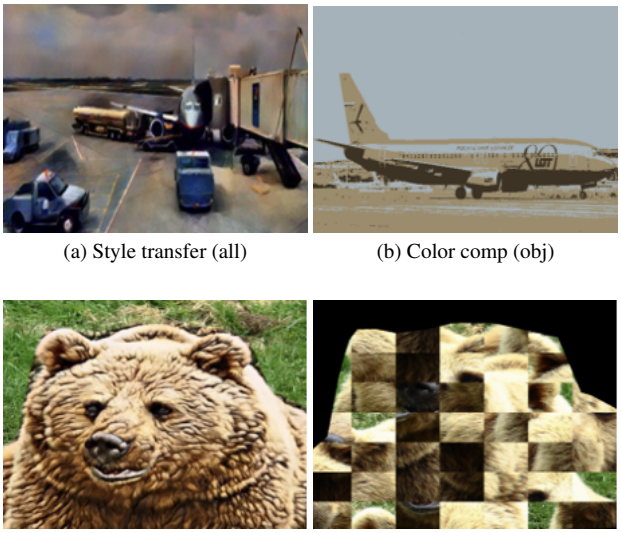
Patch shuffle image. In the previous study of image classification, the whole image was split into patches and randomly mixed to test whether the model had texture bias. In this study, the method was modified to fit the object detection task. Each bounding box was separated into several patches by obtaining the maximum common divisor of the bounding box height and width, and the overlapping area of each patch with the object mask was calculated. Only patches that overlap with the mask by a certain percentage or more, were randomly mixed within each object. The reason for this is that if the entire image is randomly mixed, objects for all classes will be mixed and the detection model wouldn't know where to draw the bounding box. To prevent this, the original bounding box was maintained, and shape information was distorted only within each object as much as possible without affecting the background. The deformation process can be found in Figure 4. This variation has obj, mask images.

For measuring texture bias we used the datasets whose backgrounds were also transformed (the 'all' dataset among 3 background variations)

Examples of all modification methods can be found in Figure 5.

4.2. Models

We chose three models - CO-DETR, DiffusionDet and YoloX[4, 10, 34] - which all extract image features using CNN but whose architectures vary greatly. CO-DETR is a modified version of DETR which uses transformers, DiffusionDet applies a diffusion model to generate correct bounding boxes, and YoloX uses convolution layers to classify and regress bounding boxes. We hypothesize that if similar results are observed in all three models, it is due to the characteristic of the detection task itself, not some specific architectural design.



(a) Style transfer (all) (b) Color comp (obj)

(c) Cue conflict (obj) (d) Patch shuffle (mask)

Figure 5. Data augmentation

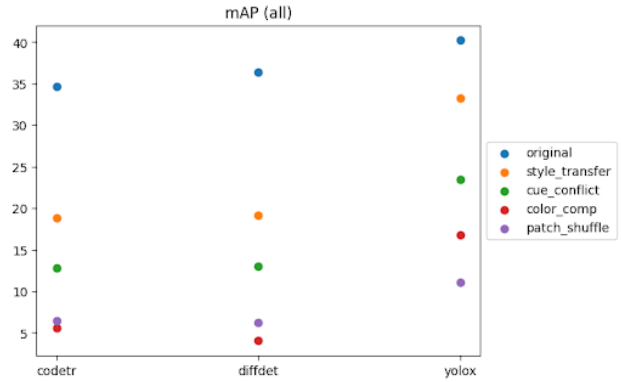


Figure 6. mAP values after data augmentation

4.3. Results

By comparing model performance in texture deformed images (style transfer, cue conflict, and color compression) with shape deformed ones (patch shuffle) we tested whether current SOTA models also have the problem of texture bias. We discovered that current detection models do not rely heavily on texture. Even for classes where texture bias of classification models was high, detection models were more prone to shape deform than texture deform. Additionally, the three models performed differently when background was removed hinting the need for further study on each model's background bias.

model	shape-biased	texture-biased
CO-DETR	0 (person), 1 (bicycle), 2 (car), 12 (parking meter), 17 (horse), 19 (cow), **20 (elephant)**, 22 (zebra), 25 (umbrella), 27 (tie), 31 (snowboard), 60 (dining table), 62 (tv), 63 (laptop), 64 (mouse), 67 (cell phone), 71 (sink)	33 (kite)
DiffusionDet	17 (horse), **20 (elephant)**, **21 (bear)**, 22 (zebra), 25 (umbrella), 27 (tie), 54 (donut), 63 (laptop), 77 (teddy bear)	8 (boat), 9 (traffic light), 33 (kite)
YoloX	0 (person), 1 (bicycle), 2 (car), 3 (motorbike), 5 (bus), 6 (train), 7 (truck), 10 (fire hydrant), 12 (parking meter), **14 (bird)**, **15 (cat)**, 16 (dog), 17 (horse), 18 (sheep), 19 (cow), **20 (elephant)**, **21 (bear)**, 22 (zebra), 23 (giraffe), 24 (backpack), 25 (umbrella), 26 (handbag), 27 (tie), 28 (suitcase), 29 (frisbee), 34 (baseball bat), 35 (baseball glove), 36 (skateboard), 37 (surfboard), 38 (tennis racket), 39 (bottle), 40 (wine glass), 41 (cup), 42 (fork), **43 (knife)**, 44 (spoon), 45 (bowl), 54 (donut), 55 (cake), 56 (chair), 58 (potted plant), 60 (dining table), 61 (toilet), 62 (tv), 63 (laptop), 64 (mouse), **66 (keyboard)**, 67 (cell phone), 68 (microwave), 69 (oven), 71 (sink), 77 (teddy bear)	33 (kite), 78 (hair drier)

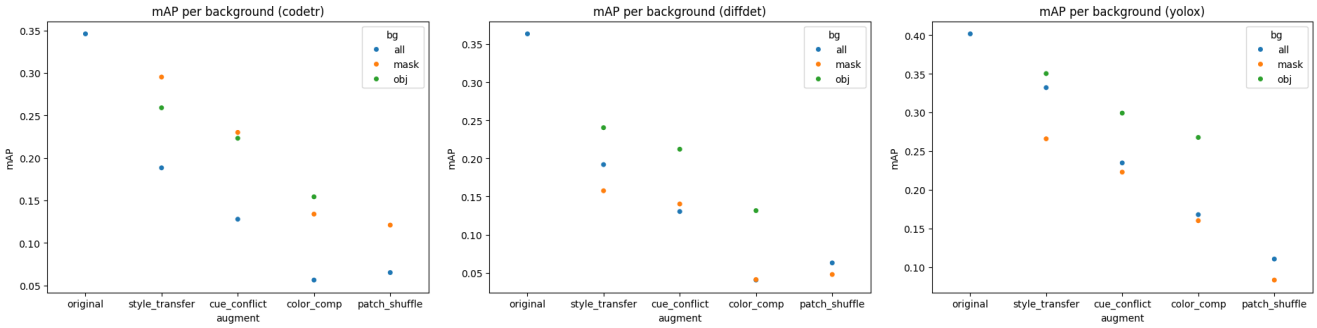
Table 1. Classes where shape/texture was important (threshold = 1)

Classes where classification models focused on texture a lot (among the 16 imageNet superclasses) are written in bold.

Low texture bias in detection models. As in Figure 6. all three models showed low mAP in patch shuffled images compared to images whose texture was more heavily deformed. This shows that object detection models do not rely on texture, unlike classification models. YoloX scored the highest mAP in all augmentation methods contrary to our belief that more recent models will fare better. All three models' mAP in the original COCO dataset were lower than the score reported by the original model's paper which might be indicative of fluctuations in test performance. More robust testing will be needed to compare corruption robustness between models. However our main finding that detection models show low texture bias is reliable based on the consistently low performance in patch shuffled data. This is aligned with our preliminary results.

Texture bias per class. We examined whether the degree of texture bias differs across classes. In the preliminary study we grouped classes based on accuracy propensity but here we took a different approach to objectively compare 80 classes. We compared whether the maximum AP value among the texture augmented datasets - style transfer, cue conflict, color compression - is higher or lower by a value of 1%p compared to patch shuffled data. If the former is higher then we interpret the model relies more on shape to detect that class. If the latter is higher we conclude the model is sensitive to texture for that specific class.

From the results displayed in Table 1, we see that only two or three classes were texture sensitive while there were 17 shape sensitive classes for CO-DETR, 9 for DiffusionDet and 52 classes for YoloX. The 'elephant' class which showed high texture bias compared to other 15 superclasses in AlexNet, VGG-16, GoogLeNet and ResNet-50 classification models[11] was included as a shape biased class in all three detection models. The 'bear' class which was also sensitive to texture in classification models was shape biased in YoloX and DiffusionDet.



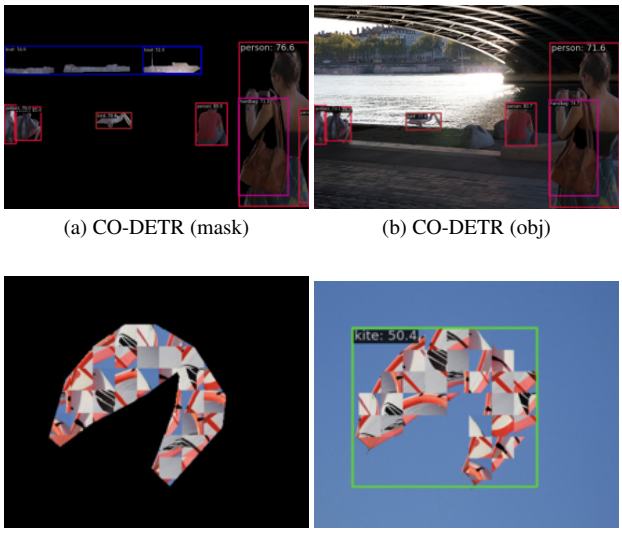
(a) CO-DETR (b) DiffusionDet (c) YoloX

Figure 7. mAP per background

Although there were more classes whose shape was more important, all three models showed high texture bias when detecting the 'kite' class. Since there were 91 instances of kite in the dataset we thought the models' inability to detect kites was meaningful and found that the kites in the dataset were usually small and their shapes were irregular. We speculate the various shapes of kites leave the model no choice but to rely on texture.

Background influence As seen in Figure 7, DiffusionDet and YoloX showed highest mAP when original background was preserved. This result is similar to YoloV.8's result in the preliminary study ('obj' dataset). This means the semantic information of the background helps them classify and detect objects.

However, CO-DETR's performance in images with the original background was lower or similar to the case where the background was blacked out ('mask' dataset). We speculate CO-DETR is less dependent on background information and removing the background actually helps the model by emphasizing the silhouette of the target object. In Figure 8 we display examples where CO-DETR detected more objects in the masked image while DiffusionDet's performance was better in the 'obj' image.



(a) CO-DETR (mask) (b) CO-DETR (obj)

(c) DiffusionDet (mask) (d) DiffusionDet (obj)

Figure 8. Background influence

4.4. Future experiment

In the course of our experiments, we found several experiments that would be meaningful. In this study, we investigated how the texture bias characteristic of CNN-based models differs in detection tasks by using the most basic version of CO-DETR with a CNN backbone. It would be beneficial to compare these findings with other high-performing models, such as those based on transformer architectures like Swin-L, to determine any significant differences. Additionally, just as humans tend to focus more on texture or shape depending on the object, it could be meaningful to analyze class-specific bias characteristics, not just the model's inherent bias.

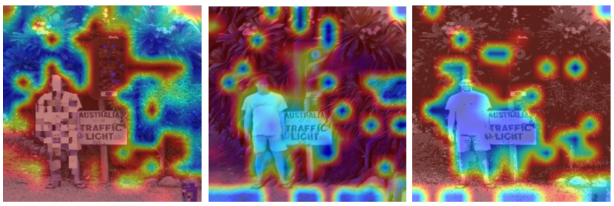
Furthermore, during our experiments, we observed distinct patterns of attention from shallow to deep layers. Analyzing these patterns could provide insights into how models interpret images. These aspects will be explored in future work.

5. Conclusion and future work

We aimed to investigate the shape/texture bias in object detection. Most detection models extract image features using a CNN-based backbone, but they also provide additional supervision for shape information to match the bounding boxes to the object's location. We expected this to result in different behavior compared to classification tasks. By improving upon existing experimental methods using our novel augmentation technique, we evaluated the perfor-

mance of more recent models. We found that while object detection models utilize features similar to those used by image classification models, they exhibit a higher shape bias due to the task of locating objects and the corresponding bounding box loss function.

Based on these findings, we concluded that the supervision provided by the dataset and the loss function corresponding to the task influence the tendency of models (e.g., texture bias or shape bias). Applying this conclusion, to focus more on shape within image classification, which is traditionally known for its texture bias, we can consider improvements from two perspectives. Firstly, the commonly used ImageNet dataset often includes multiple objects in a single image, which might hinder the model from learning object shapes. Therefore, it is necessary to crop specific objects and train on a dataset with texture deformed images. Secondly, we could introduce pseudo-labels by adding a loss function that ensures the activation weights are similar across original, style transfer, cue conflict, and color comp images, facilitating shape-focused learning. As seen in Figure 9 the attention weights in cue conflict and color comp images are similar. Since texture is deformed in these images we hypothesize the transformer based model is looking at object's shape (High attention values are in blue and low values in red). Adding a loss term to make the activated weights similar in all images with deformed texture might help the model learn what shape is. Furthermore, we hypothesize that semantic segmentation could have an even higher shape bias than object detection by utilizing silhouette information as additional supervision.



(a) Patch shuffle (b) Cue conflict (c) Color comp

(d) Patch shuffle (e) Cue conflict (f) Color comp

Figure 9. Saliency map of attention weights

In the future, if we further analyze using other methods, understanding how and why models' image perception differs from human perception, we can develop even better models and features that are not vulnerable to texture deformed data, such as paintings, resulting in even more robust performance.

References

[1] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12), 2018.

[2] Hongping Cai, Qi Wu, Tadeo Corradi, and Peter Hall. The cross-depiction problem: Computer vision algorithms for recognising objects in artwork and in photographs, 2015. arXiv preprint arXiv:1505.00110.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer International Publishing, 2020.

[4] Shoufa Chen et al. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[5] Hyunhee Chung and Kyung Ho Park. Shape prior is not all you need: Discovering balance between texture and shape bias in cnn. In *Proceedings of the Asian Conference on Computer Vision*, pages 4160–4175, 2022.

[6] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *Advances in Neural Information Processing Systems*, 2024.

[7] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015.

[8] S. Gao, Z.-Y. Li, M.-H. Yang, M.-M. Cheng, J. Han, and P. Torr. Large-scale unsupervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[10] Zheng Ge et al. Yolox: Exceeding yolo series in 2021, 2021. arXiv preprint arXiv:2107.08430.

[11] 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, 2018. arXiv abs/1811.12231.

[12] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and surface variations, 2018. arXiv preprint arXiv:1807.01697.

[13] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2020.

[14] David Kadish, Sebastian Risi, and Anders Sundnes Løvlie. Improving object detection in art images using only style transfer. In *2021 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2021.

[15] Nikolai Kalischek, Rodrigo Caye Daudt, Torben Peters, Reinhard Furrer, Jan D. Wegner, and Konrad Schindler. Biased-rigorous texture bias evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22221–22230, 2023.

[16] Guoliang Kang, Xuanyi Dong, Liang Zheng, and Yi Yang. Patchshuffle regularization, 2017. arXiv preprint arXiv:1707.07103.

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

[18] 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.

[19] 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*, pages 740–755. Springer International Publishing, 2014.

[20] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[21] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020.

[22] Chaithanya Kumar Mummadi, Ranjitha Subramaniam, Robin Hutmacher, Julien Vitay, Volker Fischer, and Jan Hendrik Metzen. Does enhanced shape bias improve neural network robustness to common corruptions?, 2021. arXiv preprint arXiv:2104.09789.

[23] OpenAI. Chatgpt (mar 14 version) [large language model], 2023.

[24] OpenAI. Dall·e 3 system card, 2023. Accessed May 13, 2024.

[25] Xinkuan Qiu, Meina Kan, Yongbin Zhou, Yanchao Bi, and Shiguang Shan. Shape-biased cnns are not always superior in out-of-distribution robustness. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2326–2335, 2024.

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015.

[28] 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, 2014. arXiv:1409.0575.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015.

[30] small yellow duck and Wendy Kan. Painter by numbers. Kaggle, 2016.

[31] Nicholas Westlake, Hongping Cai, and Peter Hall. Detecting people in artwork with cnns. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I*, pages 825–841. Springer International Publishing, 2016.

[32] Brendel Wieland and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet, 2019. arXiv preprint arXiv:1904.00760.

[33] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in neural information processing systems*, 2019.

[34] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.