Given an input image, we create two stylized images, by applying a style transfer algorithm (i.e., two different styles on the same image) such that the images share the same shape.

We feed the original and the stylized images to the same CNNs to generate latent representations.

Then we maximize the number of neurons which encode shape by computing the mutual information between the latent representations of the pair of images which share a semantic concept (i.e., shape).

The three representations are then passed through a shared classifier to obtain the classification outputs which are used to compute the cross entropy losses.

We propose a novel training loss for increasing a neural network’s ability to encode shape information.

Our shape biased models are more robust to adversarial attacks and distorted images, and generalize better to out-of-distribution examples.

We obtained all these benefits without sacrificing overall performance on ILSVRC2012 ImageNet and transfer learning on downstream tasks.

Recent works have shown that models which strongly rely on texture information to make categorical decisions perform poorly on out-of-domain examples.

Alternatively, networks which classify based on object shape have improved generalization ability and robustness.

We presented a simple and effective strategy to learn shape-centric representations for object recognition while improving the network’s robustness and generalization.

We achieved robustness without sacrificing overall classification performance.

We model showed a significant improvement in its robustness to various attacks and distortions.

We directly assess the quality of the shape-based representations by fine-tuning on different downstream tasks.

Our approach marginally outperforms the baselines across various tasks.

Evaluated on NIPS-2017-non-targeted-adversarial attack dataset.

Our method is significantly more robust than the ShapeNet.