Maximizing Mutual Shape Information

Md Amirul Islam[1,5]
cs.ryerson.ca/~amirul
Matthew Kowal[2,5]
mkowal2.github.io
Patrick Esser[7]
github.com/pesser/
Björn Ommer[4]
ommer-lab.com/people/ommer/
Konstantinos G. Derpanis[2,5,6]
www.eecs.yorku.ca/~kosta
Neil D. B. Bruce[3,5]
socs.uoguelph.ca/~brucen

[1] Ryerson University, Canada
[2] York University, Canada
[3] University of Guelph, Canada
[4] University of Munich, Germany
[5] Vector Institute for AI, Canada
[6] Samsung AI Centre Toronto
[7] Runway ML

Abstract

Recent works have shown that training a neural network with different stylized images increases the shape bias while improving robustness to common corruptions and adversarial attacks. In this work, we propose a novel training loss for increasing a neural network's ability to encode shape information. This is done by maximizing the mutual information between a network's representations of two stylized images which share the same *shape*. Compared to similar approaches, we show that our method induces a stronger inductive bias in the network towards encoding shape-based representations. Additionally, we show that our model is more robust to adversarial attacks and distorted images, and generalizes better to out-of-distribution examples. We obtain all these benefits without sacrificing overall performance on ILSVRC2012 ImageNet and transfer learning on downstream tasks (e.g., object recognition, semantic segmentation, texture recognition).

1 Introduction

Despite the impressive performance of deep neural networks (DNNs) on a variety of computer vision tasks, they are susceptible to making predictions based on spurious correlations, such as the texture within an image rather than an object's shape. Indeed, recent studies have shown convolutional neural networks (CNNs) have a 'texture bias' [14]. Given an image with conflicting cues, such as a stylized image where the object's boundary and texture have different classes, a CNN will tend towards predicting the texture label, while humans are far more likely to make predictions based on an object's shape [14] (see Fig. 1). A natural question to ask is then 'Why do CNNs focus on texture rather than object shape?' where shape is defined as the 2D silhouette of an object. One answer to this question is that ImageNet [6] (the main dataset used in these studies) is largely *solvable from texture-based representations*.

Thus, there is a value alignment problem. The network being trained is exclusively focused on minimizing the loss function, which is most easily done by using object textures. Without additional constraints to shape the loss function, the network will not learn to discriminate using object shapes which requires learning more long-range connections with a larger receptive field. This explanation is supported by research showing that CNNs with a restricted receptive field size achieve performance comparable to networks having no such restrictions [4].

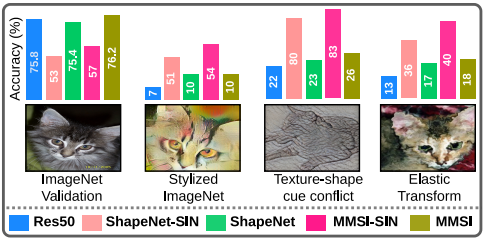

Figure 1: Previous works to induce shape-bias, such as ShapeNet [14] (i.e., a ResNet-50 [17] trained on stylized ImageNet [14] (SIN)) have been shown to make predictions based on *local* shape cues [21]. We propose a novel objective function which provides a global shape training signal by maximizing the mutual shape information (MMSI) between latent representations. We show a ResNet50 network trained with our objective outperforms ShapeNet on normal, stylized, cue-conflict, and corrupted image datasets.

Similarly, recent works have shown that models which strongly rely on *texture* information to make categorical decisions perform poorly on out-of-domain examples and are more vulnerable against different types of corruptions and adversarial attacks compared to the models which rely on *shape* information [12, 30]. Extrapolating this finding to applications where avoiding failure is critical (e.g., self-driving vehicles or medical image segmentation) suggests that current deep learning systems are too brittle, and we must encourage the network to make decisions based on more robust image features. Alternatively, networks which classify based on object shape have improved generalization ability and robustness in certain domains [18, 28]. Therefore a fruitful area of exploration is designing models which are biased towards, and encode more, shape information in their latent representations.

In this paper, we propose a novel objective function based on an approximation of mutual *shape* information between the latent representations of image pairs. Our motivation for using this particular objective is to correct the value alignment problem and encourage solutions which constrain the network to learn shape-based representations. Our objective provides a *global* signal of shape information contained in the image (i.e., in the entire latent representation). In contrast, existing works [14, 28] which train on image-level labels has been shown to make predictions based on local shape cues [21]. To show the efficacy of our approach, we first describe how to apply it in an end-to-end training pipeline for object recognition. Next, we evaluate the robustness to distortions and adversarial attacks, out-of-distribution examples, performance on ImageNet, and finally transfer learning on a variety of downstream tasks. In summary, our contributions are as follows:

- We propose a new objective function which maximizes the amount of shape information encoded in a CNN's representations. We show quantitatively that optimizing this objective increases both the shape-bias and the number of shape encoding neurons more than previous methods.
- We demonstrate that our models have substantially more robustness against distortions, adversarial attacks, and have better performance on out-of-distribution examples.
- We show empirically that our training procedure obtains these benefits *without* sacrificing performance on ImageNet as well as transfer learning to downstream tasks.

2 Related Work

Shape-Texture Bias. Visualization techniques [3, 31, 35] have been used to demonstrate that the early layers in CNNs are maximally activated for high-frequency patterns, such as textures, while deeper layers in the network activate for more abstract patterns such as object shape. Despite shape information causing high activations in the network, recent work [14] showed that existing CNN architectures trained on ImageNet [6] (such as ResNet [17] or AlexNet [25]) are biased towards making predictions based on image texture and can achieve human-level performance from solely image textures. Follow-up work [19] explored this phenomenon in much more detail with various architectures and training paradigms. Finally, [21] proposed two new shape-based metrics which they showed can accurately estimate the number of shape encoding neurons as well as quantify the amount of global shape information encoded on a *per-pixel* level.

Learning Shape-Based Representations. After pointing out the significant bias towards texture on cue-conflict images (i.e., images with both a shape and texture label), [14] proposed to train networks on stylized images to bias a CNN towards learning shape. [21] showed that networks trained with this approach fail to capture the global object shape and makes decisions based on *partial* shape. Another line of work aims to *ignore the texture* contained in an image. [34] proposed a module based on the gray-level co-occurrence matrix [16, 27] to capture textural information which is removed from the learned representations using reverse gradients [13]. [28] proposed to debias the model towards both shape and texture, by training a CNN to predict both the shape and texture label of a stylized image. Low-pass filters have also been used to debias models away from texture [32]. Given claims from existing work that shape-biased models are more robust to common distortions [14], [30] performs a systematic study of how texture-shape bias phenomenon impacts a model's robustness to corruptions and distortions. They found that robustness arises due to training on stylized images, and is not always correlated to the shape bias of the model.

3 Maximizing Mutual Shape Information

In this section, we first introduce our Maximize Mutual Shape Information (MMSI) objective function. Next, we show its differentiability with respect to the network weights. Finally, we describe the image sampling procedure required for our method and training details.

3.1 Estimating the Mutual Shape Information

Our overall approach is shown in Fig. 2. We take inspiration from recent work [8, 21, 22], where they approximate the number of neurons which encode *shape* and *texture* by estimating the mutual information between the latent representations of a pair of images which share a semantic concept (i.e., *shape* or *texture*). Using the mutual information estimate as an objective provides a way to *maximize the number of neurons which encode shape*, and thus the global shape-based signal we are interested in. The key assumption when using this approximation is that the probability distribution of the latent representations across the dataset is jointly *Gaussian*. Note that this assumption does not need to be exact for benefits to arise from our training process, as shown by our consistent improvement in performance (Sec. 4.2). With this assumption, given a pair of images, $\{I^a, I^b\}$ with latent representations,

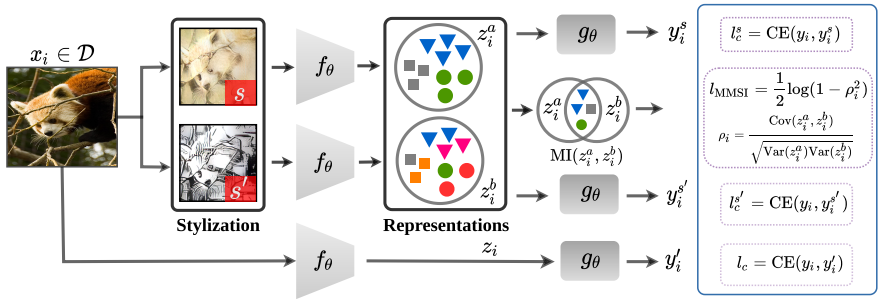

Figure 2: Illustration of our proposed approach for learning a shape-based representation by maximizing the mutual shape information (MMSI) between the latent representations of image pairs. Given an input image, $x_i \in \mathcal{D}$, we create two stylized images, $\{s, s'\}$, by applying a style transfer algorithm (i.e., two different styles on the same image) such that the images share the same *shape*. Next, we feed the original and the stylized images to the same CNNs, f_θ, to generate latent representations, z_i, z_i^a, z_i^b, respectively. We then compute the mutual shape information between two stylized representations, z_i^a, z_i^b, which is optimized as a loss function, L_{MMSI}. The three representations are then passed through a shared classifier, g_θ, to obtain the classification outputs which are used to compute the cross entropy losses.

$\{z_i^a, z_i^b\}$, which share the same semantic factor (i.e., *shape*), the correlation coefficient, ρ_i, provides a tight lower bound on the true mutual information [12, 23] as follows:

$$\text{MI}(z_i^a, z_i^b) \geq -\tfrac{1}{2}\log(1-\rho_i^2), \quad \text{where} \quad \rho_i = \frac{\text{Cov}\left(z_i^a, z_i^b\right)}{\sqrt{\text{Var}(z_i^a)\,\text{Var}(z_i^b)}}. \tag{1}$$

Given the above approximation, we simply multiply the mutual information by -1 such that we minimize it during training, resulting in our MMSI objective:

$$l_{\text{MMSI}}(z_i^a, z_i^b) = \tfrac{1}{2}\log(1-\rho_i^2). \tag{2}$$

Note that it is necessary that the lower bound on the mutual information is differentiable with respect to the network weights, Θ. We compute the derivative as follows, by dropping the sample index i for readability:

$$\begin{aligned} \nabla_\Theta\left[\text{MI}(z^a, z^b)\right] &\approx \frac{1}{2}\nabla_\Theta\left[\log(1-\rho^2)\right] \\ &= \frac{1}{2}\frac{1}{(1-\rho^2)}\cdot -2\rho\nabla_\Theta\rho \\ &= \frac{\rho}{(\rho^2-1)}\cdot\nabla_\Theta\rho. \end{aligned} \tag{3}$$

In Eq. 3, ρ is differentiable with respect to θ as ρ is simply the correlation coefficient between the two latent representations, z^a and z^b, which are the outputs of a CNN and therefore differentiable (i.e., since the covariance and variance operations are trivially differentiable).

Another perspective of the MMSI loss, is that it captures the *invariance* between the two representations z_i^a and z_i^b. Therefore, it is crucial that the image pair I^a and I^b share the same *shape*; however, the MMSI loss will also measure invariances other than *shape* in the representation space. For instance, if we choose images which share the same *shape* and *color*, the MMSI loss will capture both *shape* and *color*. Despite this, we argue and show

empirically that as long as the same *shape* is sufficiently contained between image pairs, MMSI loss will capture an adequate amount of *shape* to admit beneficial properties to the network, such as increased robustness. We next describe the image sampling procedure which ensures that the image pairs share the same shape while minimizing other invariances.

3.2 Sampling Images with Similar Shape

We now discuss the procedure of sampling image pairs which share the semantic factor: *shape*. This will ensure that object *shape* is always contained in the mutual information between the latent representations. We use a modified version of the Stylized ImageNet (SIN) [14] dataset, which uses a style transfer algorithm [20] to impose a random style, taken from a dataset of artistic paintings, onto an ImageNet image. This stylization process obfuscates the texture information of the image while leaving the shape of the object largely visible. The modification is simply that we use two styles per image in training, contrasting the original SIN dataset [14] where each image only has a single stylized counterpart.

3.3 Training the Network

We first sample a mini-batch from the dataset of normal images, $\mathcal{D}$, and stylize each image, x_i, in the batch with different styles, s and s'. We feed each stylized image through the encoder, f_θ, to obtain the representations from the last convolutional layer of the network before the classifier. Next, we use a global average pooling layer to produce latent representations, z_i^a and z_i^b, of dimensions equal to the number of channels (e.g., 2048 for ResNet50), which are then used to calculate the MMSI objective. We also feed the stylized images, along with the normal image, through the classifier, g_θ, to obtain class logits, $y_i^s, y_i^{s'}$, and y_i', respectively. We calculate the cross entropy (CE) loss between each of the class logits, and the ground-truth class label y_i. Our final loss, L, is the sum of the MMSI loss, l_{MMSI}, and the three *CE* losses, each with a corresponding weighting term, λ_j. Our final loss is given as follows:

$$L = l_c + \lambda_m \cdot l_{\text{MMSI}} + \lambda_s \cdot l_c^s + \lambda_s \cdot l_c^{s'}. \tag{4}$$

4 Experiments

Implementation Details. We empirically set the value of the loss weights, λ_m and λ_s, to 0.8 and 0.1, respectively (see Table 7 for ablation). For a fair comparison with other shape-based methods trained on ImageNet variants, we choose a similar set of hyper-parameters which are described in the supplementary materials. Following [14], we analyze the following joint training schemes. **ResNet-IN:** ResNet trained on ImageNet. **ShapeNet-SIN:** ResNet [17] trained on stylized ImageNet. **ShapeNet-SIN+IN:** ResNet variant jointly trained on ImageNet and stylized ImageNet. **ShapeNet**: The ShapeNet-SIN+IN model is further fine-tuned on ImageNet (denoted as ShapeNet in [14]). **MMSI-SIN:** The same structure as ShapeNet-SIN except our proposed MMSI loss is added. **MMSI-SIN+IN:** The same structure as ShapeNet-SIN+IN except using the MMSI loss. **MMSI**: The MMSI-SIN+IN model is further fine-tuned on ImageNet (i.e., similar to ShapeNet [14]).

4.1 Evaluation of Shape Bias and Dimensionality

Shape Bias. We compute the shape bias of CNNs trained on ImageNet which differ significantly in their ability to encode shape information. Following [14], we generate 1,280

Methods	ResNet-50				ResNet-34				ResNet-18			
	Bias (%)		Factor $\lvert z_k \rvert$ (%)		Bias (%)		Factor $\lvert z_k \rvert$ (%)		Bias (%)		Factor $\lvert z_k \rvert$ (%)	
	Shape	Texture	Shape	Texture	Shape	Texture	Shape	Texture	Shape	Texture	Shape	Texture
ShapeNet-SIN	79.8	20.2	26.2	23.3	86.6	13.4	28.1	20.7	82.1	17.9	26.4	22.1
MMSI-SIN	**83.1**	16.9	**36.0**	17.1	**87.5**	12.5	**34.2**	17.6	**86.4**	13.6	**35.9**	17.4
ShapeNet-SIN+IN	39.3	60.7	23.7	24.4	43.8	56.2	23.4	24.8	39.2	60.8	23.4	24.6
MMSI-SIN+IN	**51.9**	48.1	**32.7**	19.0	**46.6**	53.4	**30.3**	19.9	**56.3**	43.7	**31.4**	19.1
ResNet-IN	21.8	78.2	17.0	33.8	25.6	74.4	18.0	31.4	24.1	75.9	17.6	31.8
ShapeNet	22.9	77.1	18.4	31.2	25.1	74.9	18.2	31.2	26.2	73.8	**18.3**	31.3
MMSI	**26.1**	73.9	**18.8**	30.2	**26.7**	73.3	**19.1**	29.7	**27.1**	72.9	18.2	31.4

Table 1: Comparison of shape bias [14] and shape dimensionality [21] for different ResNet [17] variants. We compare our approach, maximize mutual shape information (MMSI), with vanilla ImageNet training (IN) and different training procedures of ShapeNet [14]. The dimensionality measures the percentage of neurons in the latent representation (i.e., the final stage after the global average pooling layer) which encode shape or texture.

cue-conflict images by performing style transfer based on Adaptive Instance Normalization [20] with the style transfer coefficient set to $\alpha = 0.5$. As in [14], we evaluate the models on all 1,280 images and map the ImageNet class probabilities to the corresponding 16-class-ImageNet fine-grained categories. We only consider the subset of *correctly* classified images (i.e., either *shape* or *texture* category correctly predicted) to compute the biases. The results of this comparison are presented in Table 1 under the 'Bias' heading.

Dimensionality Estimation of Shape Neurons. To quantify the number of shape neurons encoded in the latent representation of a pretrained CNN, we use the approach of [8], where the number of neurons that represent a certain semantic concept (e.g., *shape* or *texture*) is estimated. To achieve this, we must use a dataset where we can choose image pairs which share the semantic concept *shape*. Following [21], we generate the stylized PASCAL VOC12 dataset using PyTorch-AdaIN stylize algorithm [20]. Then we follow [8, 21] to calculate the number of shape and texture encoding neurons in each network.

Discussion. Table 1 compares the shape and texture *bias* and *dimensionality* of ResNet variants trained on ImageNet under different settings. As expected, MMSI based methods produce more shape encoding neurons and 'shape bias' [14] than the baselines. For both metrics, training a model solely on SIN encodes significantly more shape than the models solely trained on IN or (SIN+IN). When the (SIN+IN) pretrained model is further fine-tuned on IN, both the number of shape neurons and shape bias significantly drop; however, MMSI still shows an increase in shape over ShapeNet [14] for both metrics. The results in Table 1 show that the global signal contained in our MMSI training objective produces networks yields increases under two metrics which measure the amount of shape encoded in the model.

4.2 Robustness of Shape-based Representations.

We now evaluate the robustness of our model against adversarial attacks and explore its generalization ability to common corruptions. Our goal is to test the hypothesis that models with a greater capacity to represent shape results in increased robustness.

Influence of Shape on Adversarial Attacks. We evaluate the robustness of different shape-based models against four different adversarial attacks: Fast Gradient Sign Method

Method	Clean	FGSM [15]				PGD [29]				I-FGSM [26]				MIM [7]			
		$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=8$	$\varepsilon=16$	$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=8$	$\varepsilon=16$	$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=8$	$\varepsilon=16$	$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=8$	$\varepsilon=16$
ShapeNet-SIN	58.1	55.7	53.5	45.4	36.1	56.4	55.2	49.9	41.2	57.0	55.5	53.3	48.5	55.2	51.7	45.2	36.4
MMSI-SIN	59.9	56.8	54.4	48.2	36.9	57.7	56.3	52.5	41.8	57.6	56.7	**54.4**	**50.3**	56.7	**54.6**	**47.3**	**36.4**
ShapeNet	91.9	51.3	26.3	8.2	2.3	53.9	37.7	29.4	22.4	50.8	34.1	24.8	13.5	37.2	11.4	2.9	0.7
MMSI	**92.5**	**73.4**	**61.2**	**49.1**	**37.7**	**77.5**	**69.5**	**57.5**	**42.5**	**78.5**	**66.3**	47.0	25.2	**68.4**	47.3	23.1	10.9

Table 2: Classification accuracy on adversarial perturbed images, where a greater ε means a larger attack. Our MMSI loss produces significantly more robustness to all attacks compared with the training protocol from ShapeNet [14].

Method	FGSM [15]				PGD [29]				I-FGSM [26]				MIM [7]			
	$\varepsilon=.2$	$\varepsilon=.3$	$\varepsilon=.4$	$\varepsilon=.5$	$\varepsilon=.2$	$\varepsilon=.3$	$\varepsilon=.4$	$\varepsilon=.5$	$\varepsilon=.2$	$\varepsilon=.3$	$\varepsilon=.4$	$\varepsilon=.5$	$\varepsilon=.2$	$\varepsilon=.3$	$\varepsilon=.4$	$\varepsilon=.5$
ResNet50-IN	53.4	37.8	30.0	24.1	39.2	23.1	12.5	6.7	43.5	25.8	**16.5**	8.8	46.1	28.9	19.5	11.1
ShapeNet-SIN	25.3	18.3	12.5	10.5	20.0	11.8	7.5	4.3	21.7	13.2	9.1	5.6	22.4	14.7	9.8	6.4
ShapeNet [14]	**54.8**	38.6	29.5	22.7	38.9	20.8	11.4	6.7	43.0	24.3	13.7	8.8	45.7	27.4	15.8	10.6
MMSI-SIN	25.2	19.3	11.5	10.4	18.4	10.2	5.4	3.0	19.6	11	6.8	3.6	21.2	12.3	7.7	4.7
MMSI	53.7	**40.3**	**31.1**	**25.0**	**41.0**	**24.3**	**13.3**	**7.4**	**43.9**	**28.2**	16.2	**9.3**	**46.5**	**30.2**	**19.8**	**12.0**

Table 3: Classification accuracy on adversarial perturbed images, where a greater ε means a larger attack. The maximize mutual shape information (MMSI) loss produces more robustness to all attacks compared with the training protocol from ShapeNet [14].

(FGSM) [15], Projected Gradient Descent (PGD) [29], Iterative FGSM (I-FGSM) [26], and Momentum Iterative fast gradient sign Method (MIM) [7] (the latter two being *iterative* attacks). We use the ImageNet compatible NeurIPS 2017 adversarial competition dataset[1] to perform all the robustness experiments. We generate adversarial images by targeting the ResNet-50 ImageNet pretrained model and evaluate on other shape-based models. The attacks are therefore *transferred* and non-targeted to the shape-based models. We choose this evaluation scheme to solely evaluate the performance difference between our model and that of other shape-based training procedures (i.e., ShapeNet [14]). We set the perturbation, ε, between 2 and 16 in all experiments with pixel values in the range 0 and 255 which restricts the maximum perturbation change per-pixel to $\varepsilon = 16/255$. We set the number of iterations of I-FGSM and MIM to 10 and 5, respectively. We report the top-1 accuracy to show the robustness of each model under different attacks.

The results of this comparison are presented in Table 2 and show that MMSI-trained models are significantly more robust to all the adversarial attacks when trained under the same data settings. Note that the difference is higher for iterative attacks (e.g., I-FGSM and MIM in Table 2, right). Interestingly, our MMSI method, which has less shape bias and shape encoding neurons than MMSI-SIN, is more robust against attacks compared to MMSI-SIN when the value of ε is relatively small; however, for higher ε values (e.g., $\varepsilon = 8, 16$) and iterative attacks, MMSI-SIN significantly outperforms the MMSI model in terms of top-1 accuracy. Apart from the iterative attacks with large ε values, MMSI has the highest robustness to most adversarial attacks.

Now we evaluate the robustness of shape-based models against these attacks in the *targeted* setting (i.e., each model is independently attacked). We set the perturbation, ε, between $[0.1, 0.5]$ among all experiments with pixel values in $[0, 255]$. Note that we choose smaller epsilon values for this experiment as with higher epsolon values the performance degrades too significantly to achieve any reasonable performance. Similar to *untargeted* settings, we set the iteration size of I-FGSM and MIM to 10 and 5, respectively. We report the top-1 accuracy

[1] https://www.kaggle.com/c/nips-2017-non-targeted-adversarial-attack/overview

to evaluate the robustness of each model. With respect to the shape-robustness hypothesis, we draw similar conclusions as in [30]: while shape-centric models are generally more robust, a higher shape bias does not always correlate with higher robustness against adversarial attacks, contradicting the hypothesis that a higher shape bias will necessarily increase robustness. We conclude that the robustness depends both on the type of encoding learned (i.e., shape and texture) but also the type of attack (e.g., iterative).

Influence of Shape on Common Distortions. Next, we evaluate different shape-based models in terms of their generalizability and robustness to common distortions and corruptions. For these experiments, we use the ImageNet-C [18], Stylized ImageNet [14], and ImageNet-Sketch [33] datasets. Table 4 presents the accuracy on these three datasets. MMSI-SIN outperforms ShapeNet-SIN training by a reasonable margin on the datasets: 36.4% vs. 34.9% on ImageNet-C, 54.2% vs. 50.6% on Stylized ImageNet and 27% vs. 28.1% on ImageNet-Sketch. Collectively, these results suggest that our *MMSI* training is an effective way to learn robust shape-based representations compared to the vanilla SIN training and successfully leads to consistent and substantial improvements when evaluated on shape-centric datasets. Please see the supplementary for additional analysis in ImageNet-C dataset and effects.

Method	ImageNet-C	Stylized-IN	IN-Sketch
ShapeNet-SIN	34.9	50.6	27.0
MMSI-SIN	**36.4**	**54.2**	**28.1**
ResNet-50	37.7	7.1	23.1
ShapeNet	40.1	**9.8**	**25.8**
MMSI	**40.2**	9.6	25.3

Table 4: Comparison results on three out-of-domain image recognition datasets: ImageNet-C [18], Stylized ImageNet (Stylized-IN) [14], and ImageNet-Sketch (IN-Sketch) [33].

4.3 Accuracy of Shape-based Representations

It is important to determine whether improvements in robustness and generalization ability are at the cost of performance on unaltered images. To this end, we evaluate the classification accuracy of MMSI on the ImageNet ILSVRC2012 dataset [6], the results of which are presented in Table 5. Surprisingly, MMSI does not sacrifice performance and even gains marginal improvements over the vanilla IN and ShapeNet in terms of top-1 and top-5 accuracy. Our final model MMSI improves the performance over the baseline ResNet50 by 0.4%, achieving 76.2% top-1 accuracy. These results suggest that the MMSI guides the network to learn a beneficial holistic representation of object shape, without removing the capacity to learn texture cues which are crucial to classifying the fine-grained classes. As expected, the models solely trained on SIN have lower performance. Also note that since the only difference between our objective functions and ShapeNet is the MMSI loss, the improvements over ShapeNet can be largely attributed to this loss and further demonstrate the efficacy of the MMSI objective.

Method	Fine-Tune	ResNet-50		ResNet-34	
		Top-1	Top-5	Top-1	Top-5
ShapeNet-SIN	-	53.2	77.0	54.1	77.1
MMSI-SIN	-	**56.8**	**89.9**	**55.6**	**78.8**
ResNet-IN	-	75.8	92.7	73.3	91.3
ShapeNet	IN	75.4	92.5	73.0	91.1
MMSI	IN	**76.2**	**92.9**	**73.4**	**91.3**

Table 5: Performance comparison of networks biased towards shape-based representations on the ImageNet val set.

4.4 Adaptability of Learned Representations

We now directly assess the quality of the representations learned using MMSI by fine-tuning the trained network on different tasks. We aim to see whether shape-based representations are

Method	Image Classification				Semantic Segmentation		CAMs
	VOC2007 mAP(%)	Caltech-101 Top-1 (%)	CIFAR100 Top-1 (%)	DTD Top-1 (%)	VOC12 FT	VOC12 Freeze	VOC12 mIoU (%)
ShapeNet-SIN	90.7	**91.0**	80.8	**67.0**	55.7	40.0	42.6
MMSI-SIN	**91.3**	90.0	**81.1**	**67.0**	**56.2**	**40.8**	**44.6**
ResNet-IN	93.9	94.3	82.6	68.1	62.7	49.6	48.3
ShapeNet	93.8	94.2	82.7	67.6	62.5	47.4	48.3
MMSI	**94.3**	**95.0**	**82.8**	**68.7**	**62.8**	**50.0**	**48.9**

Table 6: We evaluate the fine-tuning performance of networks biased towards shape on various downstream tasks. While the improvements are modest, our maximize mutual shape information (MMSI) model consistently achieves the best performance, demonstrating the generalizability of the learned shape-based representations from the MMSI objective.

beneficial for a number of image-level and per-pixel objectives for different datasets.

Transferability to other classification tasks. We first evaluate MMSI's shape-based representations by fine-tuning our pre-trained model on four different classification datasets: PASCAL VOC 2007 [10], Caltech-101 [11], CIFAR-100 [24], and DTD [5]. We follow the prescribed evaluation protocols and report the fine-tuning results using the standard metrics for each benchmark in Table 6. The results show that MMSI representations consistently provide a good initialization strategy to match or improve the performance on other classification datasets without any architectural modifications.

Transferability to semantic segmentation. We now assess whether MMSI's representations can generalize beyond image classification tasks. We first fine-tune MMSI on the PASCAL VOC12 dataset [9] for the task of semantic segmentation (see the supplementary for details). The comparison results with the vanilla ResNet-50 and other shape-based methods are shown in Table 6. MMSI (62.8%) marginally outperforms both ShapeNet (62.5%) and the ResNet-50-IN baseline (62.7%). We also evaluate the read-out performance on PASCAL VOC12 (shown under VOC12 *Freeze*) to test the representations without fine-tuning the weights of the network. In this setting, we freeze the weights of the ResNet-50 model and train a 1-layer convolutional layer to predict the semantic segmentation output. As expected, we see a large performance gain compared to ShapeNet, with a 2.6% improvement, and a moderate 0.4% improvement over the vanilla IN trained model.

Transfer to pseudo-label generation using CAMs. We also evaluate the ability of MMSI representations to generate pseudo-labels using Class Activation Maps (CAMs) [35]. Following [1, 2], we first generate CAMs for VOC12 training images by using a multi-label classification network [17]. For a fair comparison, we initialize each network from the IN, ShapeNet, or MMSI weights and then fine-tune each network on VOC12. We generate pseudo-labels from the raw CAMs by thresholding their confidence scores (threshold is set to 0.15) for each semantic category at every pixel. Table 6 (right column) presents the comparison results of different shape-based methods in terms of the quality of the generated pseudo-labels. We consider the generated semantic pseudo-labels as predictions and calculate the mIoU between the pseudo-label and the segmentation ground-truth from VOC12. As shown in Table 6, MMSI marginally outperforms the ShapeNet [14] and IN baselines. As the CAM generations rely on both shape and texture cues, these results further demonstrate the ability of the MMSI to encourage learning global object shape and texture information.

4.5 Analysis of the MMSI Objective

Method	mAP(%)
ResNet-50 [17]	91.8
+ MMSI ($\lambda_m = 0.8$)	92.5
+ MMSI ($\lambda_m = 0.8$) + style loss ($\lambda_s = 0.1$)	**93.4**
+ $\text{MMSI}_{\text{S5+S4}}$ + style loss ($\lambda_s = 0.1$)	93.2
+ $\text{MMSI}_{\text{S5+S4+S3+S2}}$ + style loss ($\lambda_s = 0.1$)	93.3

Table 7: Ablation study to test the hyper-parameters of MMSI objective function. We evaluate the multi-label classification performance on VOC12 [9] to compare training objectives in ResNet-50 model.

We now provide an ablation study under several different settings to investigate the MMSI objective and motivate the selected hyperparameters. We focus on two major components to validate the significance of MMSI: (i) training a network with and w/o MMSI objective and (ii) applying MMSI on different stages of a network. We present quantitative results comparing different settings in Table 7. We choose the PASCAL VOC12 [9] classification dataset in this experiment for computational efficiency. The classification performance, mAP, is improved (91.8% vs. 92.5%) when the MMSI with a loss weight of $\lambda_m = 0.8$ is added with the baseline ResNet-50 [17], which can be further improved by including cross-entropy (CE) losses on the stylized images and non-stylized image. We empirically set the loss weight, $\lambda_s = 0.1$, for the CE on the stylized images as increasing λ_s degrades the performance. We also tried applying the MMSI loss at different stages of the network; however, the overall accuracy is not affected.

5 Discussion and Conclusion

We presented a simple and effective strategy to learn shape-centric representations for object recognition while improving the network's robustness and generalization. Our MMSI loss maximizes the mutual information between latent representations of image pairs sharing the same *shape*. We argued that this objective provides a more robust *global* shape training signal, contrasting previous approaches which simply use image-level labels with stylized images [14, 28]. We demonstrated the validity of this claim by using the previous metrics of quantifying shape encoding, i.e., shape bias [14] and the number of shape encoding neurons [21]. As expected, our approach contains significantly more shape in both regards.

While the main goals of this approach are to improve generalization and robustness, it is important to also evaluate the performance on unaltered images using the MMSI objective. Interestingly, we showed marginal improvements on ImageNet [6] both over the standard training procedure and ShapeNet [14]. We explored our model's robustness to adversarial attacks and various distortions, as well as generalizabilty to downstream tasks. As expected, our model showed a significant improvement in its robustness to various attacks and distortions. We also concluded that iterative attacks (e.g., MIM [7]) harm the SIN trained models more than the fine-tuned models, supporting previous work [30]. Finally, we showed that our MMSI model also provides a strong pre-training initialization for a handful of different classification and segmentation tasks, as well as showed moderate improvements when generating CAMs based pseudo-labels. It is clear that the shape-centric representations learned using the MMSI loss are beneficial in a multitude of ways. On top of improving robustness, generalizability, and pre-training initialization for downstream tasks, we believe this simple yet effective scheme to learn global object shape from 2D images will encourage the community to explore similar research avenues to make inference models more robust and better overall.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018.

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019.

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017.

[4] Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In *ICLR*, 2019.

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018.

[8] Patrick Esser, Robin Rombach, and Björn Ommer. A disentangling invertible interpretation network for explaining latent representations. In *CVPR*, 2020.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html, 2010.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 2010.

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *CVPRW*, 2004.

[12] David Foster and Peter Grassberger. Lower bounds on mutual information. *Physical review. E, Statistical, Nonlinear, and Soft Matter Physics*, 2011.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.

[14] 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. In *ICLR*, 2018.

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[16] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *TSMC*, 1973.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[18] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2018.

[19] Katherine L Hermann and Simon Kornblith. Exploring the origins and prevalence of texture bias in convolutional neural networks. In *NeurIPS*, 2020.

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[21] Md Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Björn Ommer, Konstantinos G. Derpanis, and Neil Bruce. Shape or texture: Understanding discriminative features in CNNs. In *ICLR*, 2021.

[22] Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil DB Bruce, Richard P Wildes, and Konstantinos G Derpanis. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *CVPR*, 2022.

[23] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, 2004.

[24] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[26] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLRW*, 2017.

[27] SWC Lam. Texture feature extraction using gray level gradient based co-occurence matrices. In *ICMSC*, 1996.

[28] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. In *ICLR*, 2021.

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[30] 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? In *ICLR*, 2021.

[31] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017.

[32] Samarth Sinha, Animesh Garg, and Hugo Larochelle. Curriculum by smoothing. In *NeurIPS*, 2020.

[33] Haohan Wang, Songwei Ge, Eric P Xing, and Zachary C Lipton. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

[34] Haohan Wang, Zexue He, and Eric P. Xing. Learning robust representations by projecting superficial statistics out. In *ICLR*, 2019.

[35] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.