

SUPPRESSING TEXTURE TO BUILD REPRESENTATIONS FOR BETTER TRANSFER LEARNING

Anonymous authors

Paper under double-blind review

1 TRAINING WITH STYLIZED IMAGENET

Figure 2 shows the training curves for both Stylized ImageNet and Standard ImageNet. We see that the model quickly saturates when using Stylized ImageNet. This leads to a low performance on downstream tasks. Our hypothesis is that the model rapidly learns to exploit some regularities in the texture introduced by the GANs to easily solve the MoCoV2 and Jigsaw tasks. This means that the self-supervised model has the tendency to take shortcuts in the presence of regular textures. The aim of our paper has been to investigate such shortcomings and provide appropriate solutions to the issues.

2 JIGSAW TASK

In addition to MoCoV2 we also show results with Jigsaw pretext task.

2.1 METHOD

The goal in Jigsaw is to infer correct ordering of the given regions from an image. Following Noroozi & Favaro (2016), the typical setting is to divide an image into nine non-overlapping square patches and randomly shuffle them. A CNN is trained to predict the original permutation of these patches in the image. Jigsaw++ (Noroozi et al., 2018) extended this idea and replaced some patches with random patches. These patch based methods come with their own issues and there has been some recent effort to solve them. Mundhenk et al. (2017) describe an easy short-cut CNNs take that utilizes the chromatic aberration produced due to different wavelengths exiting the lens at different angles. The authors provided a solution to the problem of chromatic aberration by removing cues from the images and also allowing the color pattern to be partially preserved. They also address the problem of true spatial extent that network sees in patch based methods by yoking the patch jitter to create a random crop effect.

2.2 RESULTS USING JIGSAW PRETRAINING

The baseline model in the case of Jigsaw is pre-trained using the standard ImageNet dataset. Unlike Noroozi & Favaro (2016), we use ResNet18 as our backbone instead of Alexnet (Krizhevsky et al., 2017) to take advantage of deeper layers and capture better image representations. We obtained these results by building on top of a publicly available implementation¹. Table 1 shows results for Jigsaw models trained and tested on different datasets. We observe that the Jigsaw model trained on the Cartoon dataset outperforms the baseline methods by 2.52 mAP and Anisotropic ImageNet outperforms the baseline methods by 1.8 mAP on the PASCAL VOC image classification dataset. On object detection Bilateral ImageNet outperforms the baseline Jigsaw model by 0.78 mAP. On semantic segmentation Anisotropic Imagenet outperforms the baseline Jigsaw models by 8.1 mAP. Traditionally semantic segmentation has been a difficult task for Self-Supervised methods (Noroozi & Favaro, 2016; Caron et al., 2018) and improvement of this order on semantic segmentation shows the effectiveness of removing texture.

We also show results on ImageNet classification as the downstream task. Due to its large scale, it is usually infeasible to fine-tune the whole network for the final task every time. Therefore, following prior work (Caron et al., 2018), we only fine-tune a linear classifier. The inputs to this classifier are

¹<https://github.com/bbrattoli/JigsawPuzzlePytorch>

Table 1: Comparison of our approach with Jigsaw baseline methods. Using our best model, we improve 2.52 mAP in VOC classification , 0.78 mAP on VOC detection and 8.1 mAP on VOC semantic segmentation(SS) over the baseline models. Note that Stylized ImageNet performs poorly on VOC classification due to the visual shortcuts.

Method	Dataset Size	VOC Cls.	VOC Det.	SS
Baseline	1.2M	74.82	61.98	27.1
Stylized (Geirhos et al., 2018)	1.2M	13.81	28.13	10.12
Gaussian ImageNet	2×1.2M	75.49	62.39	27.9
Bilateral ImageNet	2×1.2M	74.55	62.74	28.9
Only Anisotropic	1.2M	74.52	61.85	32.7
Anisotropic ImageNet	2×1.2M	76.77	61.59	35.2
Cartoon ImageNet	2×1.2M	77.34	59.31	34.1

Table 2: ImageNet classification by finetuning the last FC layer. Features from the conv layers are kept unchanged. This experiment helps evaluate the quality of features learnt by the convolutional layers.

Method	Dataset Size	VOC Cls	VOC Det.	ImageNet Cls. Acc
Jigsaw Baseline	1.2M	74.82	61.98	26.17
Jigsaw anisotropic	2×1.2M	76.77	61.59	26.67

the features from a convolution layer in the network. Note that while fine-tuning for the final task, we keep the backbone frozen. Therefore, the performance of the linear classifier can be seen as a direct representation of the quality of the features obtained from the CNN. We report the results of this experiment on ImageNet in Table 2. Adding the Anisotropic ImageNet dataset to this model gives a further improvement of 0.5%.

2.3 JIGSAW USING ALEXNET AS BACKBONE

Our improvement when using Anisotropic ImageNet is not restricted to the backbone. Traditionally in Self-Supervised learning one of the most followed architectures is Alexnet (Noroozi & Favaro, 2016; Doersch et al., 2015; Caron et al., 2018). Following these methods, we also show results on Alexnet backbone. In Table. 3 we show results on VOC Classification when using Alexnet as the backbone. We obtain an improvement of 0.67 mAP over the baseline.

2.4 PATCH-WISE ANISOTROPIC DIFFUSION

In our best performing model, we considered all the patches for the jigsaw task to either come from the standard ImageNet or anisotropic diffusion filtered ImageNet. What if each of the 9 patches for the Jigsaw task could be either a standard patch or filtered patch? For this experiment we randomly choose a patch from the standard dataset or the filtered dataset, with equal probability. This is a much more extreme form of data augmentation and considerably increases the difficulty of the task. We got an improvement of 0.6 mAP over the baseline model for the classification task. However this is 1.1 mAP lower than the doing Anisotropic Diffusion on whole image.

Table 3: Experiments with Alexnet as the backbone. Ideas of anisotropic diffusion filter can extend to other architectures like Alexnet. The Anisotropic ImageNet model improves over the baseline by 0.67 mAP

Method	VOC 2007 Classification
Jigsaw Baseline(Our Implementation)	65.21
Jigsaw anisotropic	65.88

3 ANISTROPIC IMAGES

We show some more examples of Anistropic images obtained by applying Anisotropic diffusion filters to images from ImageNet (Deng et al., 2009) in figures 5 and 6. Notice how the images lose texture information. This makes it more difficult for models to find shortcuts. This, in turn, leads to better semantic representations learned by the model which leads to higher performance on downstream tasks.

4 OTHER TEXTURE REMOVING METHODS

In this section we give details of other texture removing methods.

4.1 BILATERAL FILTERING

Bilateral Filtering (Tomasi & Manduchi, 1998) is an efficient method of anisotropic diffusion. In Gaussian filtering, each pixel is replaced by an average of neighboring pixels, weighted by their spatial distance. Bilateral Filtering is its extension in which weights also depend on photometric distance. This also limits smoothing across edges, in which nearby pixels have quite different intensities.

4.2 CARTOONIZATION

A more extreme method of limiting texture is to create cartoon images. To convert an image into a cartoonish image we first apply bilateral filtering to reduce the color palette of the image. Then in the second step we convert the actual image to grayscale and apply a median filter to reduce noise in the grayscale image. After this we create an edge mask from the greyscale image using adaptive thresholding. Finally we combine these two images to produce cartoonish looking images (see Fig. 2).

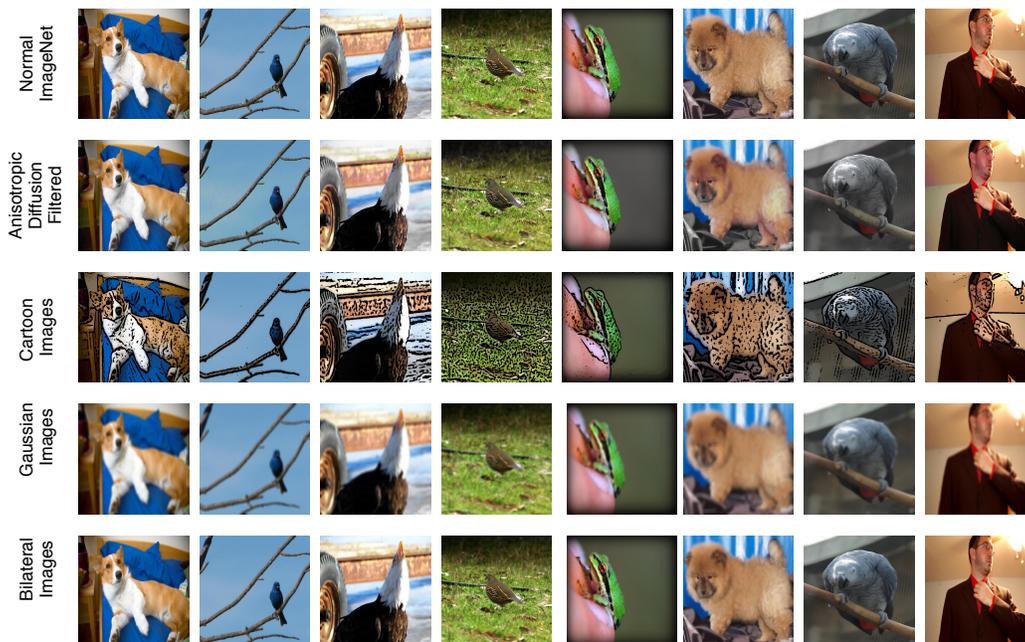


Figure 1: The figure shows examples of the effect of anisotropic diffusion. The texture on the towel in the first figure and that on the leaves and the beetle is smoothed out. This effect helps in forcing the network to rely on higher level information rather than textures.

Table 4: Results on Label corruption task. We can see that our model consistently outperforms the baseline with larger improvement upon increasing the corruption probability.

Corruption probability	Gold Fraction	Baseline(Error)	Anistropic Model(Error)
0.2	0.05	29.61	28.75
0.2	0.1	28.31	27.84
0.4	0.05	31.92	30.5
0.4	0.1	32.59	31.48
0.6	0.05	39.04	38.52
0.6	0.1	36.75	34.98
0.8	0.05	54.23	52.6
0.8	0.1	44.31	43
1.0	0.05	75.05	71.21
1.0	0.1	51.19	45.51

5 SALIENCY MAPS

5.1 SKETCH-IMAGENET SALIENCY MAPS

In Fig 7 we show some of saliency maps for Sketch-ImageNet images. We can see from saliency maps that Anistropic ImageNet has broader saliency map and has better coverage of the object as compared to ImageNet model.

5.2 SALIENCY MAPS FOR ANISTROPIC IMAGENET AND STANDARD IMAGENET MODELS

We also show some additional saliency maps in Figure 8, Figure 10, Figure 9 and Figure 11 corresponding to both the models. We can see from the figures that Anistropic ImageNet has in general diffused saliency maps.

6 IMPLEMENTATION DETAILS

Training Details. With image classification as the downstream task, we train our network for 90,000 iterations with an initial learning rate of 0.003 following (Caron et al., 2018).

For object detection we report our results for Faster-RCNN (Ren et al., 2015) using our pre-trained model as backbone. We tune hyper-parameters using the validation set. For object detection, we follow the details of (Ren et al., 2015) to train a model; 10 epochs with an initial learning rate of 0.001.

For semantic segmentation, we report our results on FCN (Shelhamer et al., 2017) using our pre-trained model as backbone. We train the FCN model for 30 epochs using an initial learning rate of 0.01.

ImageNet. We use ImageNet for all training and evaluation of image classification accuracy. For self-supervised learning, we follow (Caron et al., 2018; He et al., 2019); we train linear classifiers using features obtained from the final Residual block by freezing all convolutional layers. The performance of these linear classifiers is meant to evaluate the quality of the feature representations learnt by the convolutional layers, since the backbone is completely frozen and only fully-connected layers are being trained. We chose hyper-parameters using the validation set and report performance on the ImageNet validation set.

Note that since we use ImageNet to pre-train for self-supervised learning, there is no domain difference when we conduct inference on ImageNet, but with VOC there is. With the VOC results, we validate that the gain by our method is particularly large when there is domain shift.

Jigsaw task. In Jigsaw (Noroozi & Favaro, 2016) the image is divided into 9 non-overlapping square patches. We select 1,000 from the 9! possible permutations. All of our primary experiments on Jigsaw use ResNet18 as the backbone (He et al., 2015). We train the Jigsaw task for 90

Table 5: Comparison between Stylized ImageNet and our Anisotropic ImageNet. Following (Geirhos et al., 2018), we use ResNet50 as our backbone. We finetune our models on only the ImageNet dataset. We can see that on ImageNet classification and object detection, Anisotropic ImageNet and Stylized ImageNet have very similar performance.

Method	Finetune	Top-1 Accuracy	Top-5 Accuracy	OBJ Detection
Stylized Imagenet	-	74.59	92.14	70.6
Stylized Imagenet	IN	76.72	93.28	75.1
Anisotropic Imagenet	-	68.38	87.19	-
Anisotropic Imagenet	IN	76.71	93.26	74.27
Cartoon Imagenet	IN	76.22	93.12	72.31

Table 6: Experiments discussing the confidence and entropy of Anisotropic ImageNet and Standard ImageNet

Method	Entropy	Mean Highest probability
Anisotropic ImageNet	0.81	0.93
Standard ImageNet	1.88	0.59

epochs, with an initial learning rate of 0.01. The learning rate is reduced by a factor of 0.1 after (30, 30, 20, 10) epochs. We use the same data augmentation as in (Noroozi & Favaro, 2016). In MoCo (He et al., 2015), we use ResNet50(He et al., 2015) as the backbone, following the same procedure as mentioned in (He et al., 2019).

PASCAL VOC. Following (Caron et al., 2018) and (Chen et al., 2020), we evaluate image classification and object detection on the PASCAL VOC dataset (Everingham et al., 2009). It contains about 5,000 images in the train-val set belonging to 20 classes. Note that the image classification task is multi-label. Therefore, the metric used for evaluating both image classification and object detection is the mean Average Precision (mAP).

Training Details for Object detection for COCO based metrics: We report object detection results on (Ren et al., 2015) C4 backbone which is finetuned end to end on VOC07+12 trainval dataset and evaluated on the VOC 07 test set using the COCO suite of metrics.

Other Details. We use 4 Nvidia GTX 1080 Ti for all experiments. Pretraining on Jigsaw takes 3 days on the standard ImageNet dataset. The SGD optimizer with momentum was used for all experiments with momentum of 0.9 and weight decay of 5×10^{-4} . Cross-entropy loss was used for all experiments, mini-batch size was set to 256. Pretraining on MoCoV2 takes 6 days on 4 Nvidia P100 machines. We set all other hyperparameters following Chen et al. (2020).

7 LABEL CORRUPTION TASK

We also show results on the label corruption task in Table 4. In this task we use CIFAR100 as our dataset and we augment the CIFAR100 dataset with Anisotropic diffused images. We create a dataset double the size of original CIFAR100 dataset and use it for the task of Label Corruption (Hendrycks et al., 2019). We can see from the results that as we consistently have improvements compared to baseline. With increase in corruption probability, our results improve even more which shows that focussing on higher level features also improve accuracy in label corruption task as well.

8 CONFIDENCE OF MODELS

In this section we compare the confidence and entropy of Anisotropic Model and ImageNet model when both the models have given correct predictions. To find confidence, we generate the probability scores of correct class. After this we calculated the mean of correct probability scores on both the models. As we can see from Table 6 that Anisotropic ImageNet has larger mean which means that

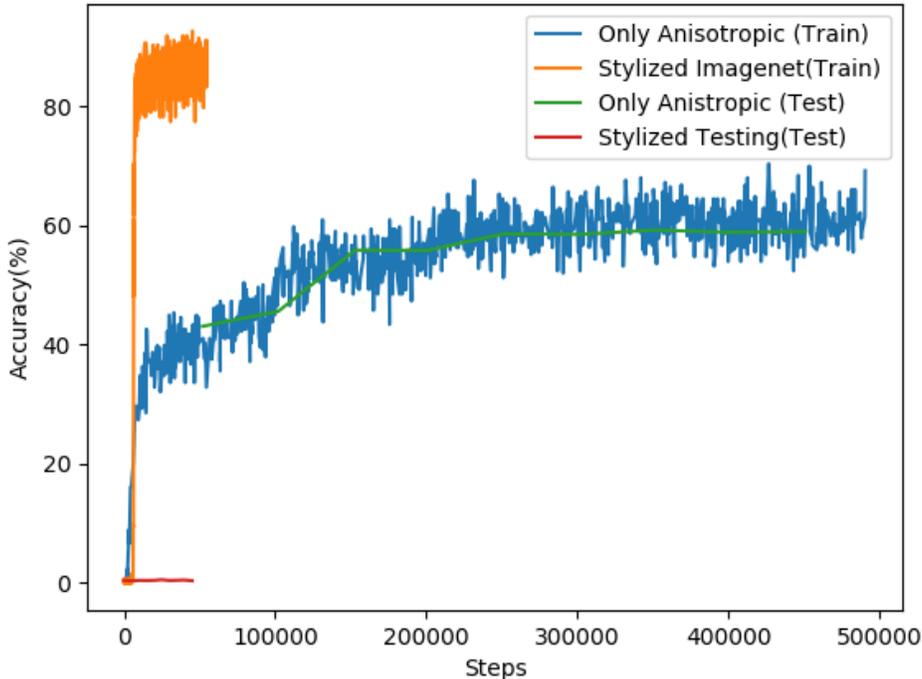


Figure 2: Training plots for the both Stylized ImageNet and Standard ImageNet. Both of these models used ResNet18 as the backbone. The plot shows that the model trained on Stylized ImageNet quickly overfits by finding shortcuts after around 6000 steps. Therefore, it gives poor performance on downstream tasks by relying on texture based shortcuts. **Refer to the Table 1 from the main paper.**

Anisotropic ImageNet has better confidence as compared to Standard ImageNet. We also calculate the entropy of output probability distribution from both the models. We can see from Table 6 Anisotropic ImageNet has lower entropy scores as compared to Standard ImageNet.

9 SALIENCY MAPS

In Fig. 3 we show the saliency maps produced by networks trained using the combined dataset and the original ImageNet dataset. We use GradCam(Selvaraju et al., 2016) to calculate the saliency maps. We can see that Anisotropic ImageNet has saliency maps that spread out over a bigger area and that include the outlines of the objects. This suggests that it attends less to texture and more to overall holistic shape. In contrast, ImageNet trained models have narrower saliency maps that miss the overall shape and focus on localized regions, suggesting an attention to texture. In Fig. 3(f-j) we show these for the case where the Anisotropic model gives the correct prediction and the ImageNet model fails. For example in Fig. 3(j), we see that the network trained on ImageNet alone is not focusing on the whole bird and is only focusing on the body to make the decision whereas the one trained with Anisotropic ImageNet is focusing on complete bird to make a decision. We see a similar trend in the cases where both the models give the correct prediction (Fig. 3(a-e)). In the case where Anisotropic model makes incorrect predictions and ImageNet model (Fig. 3(k-o)) is correct we see the saliency maps are still diffused, but we fail to capture the whole object leading to incorrect predictions.

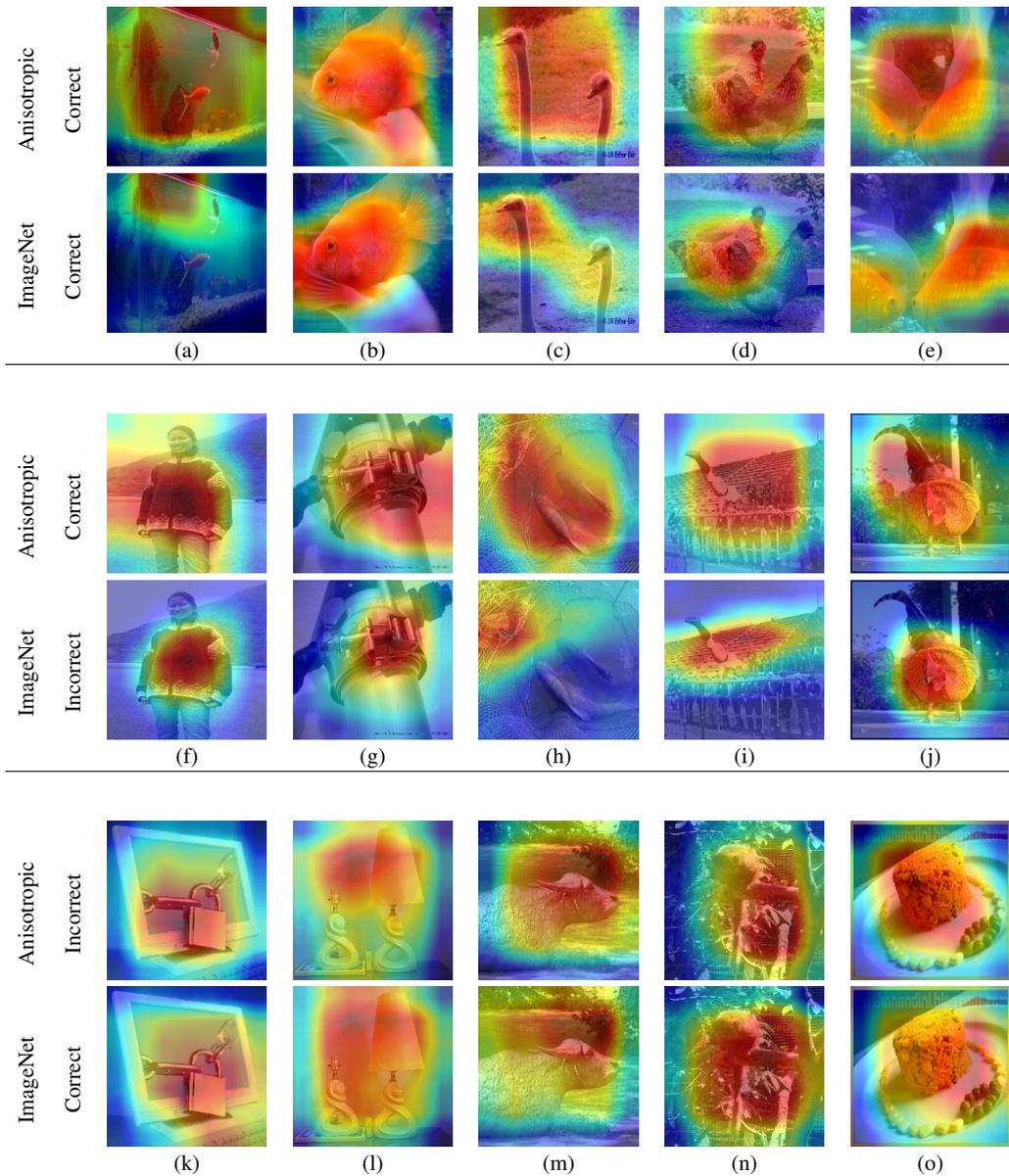


Figure 3: Saliency maps on three different set of images. The text on the left of the row indicates whether Anisotropic model or ImageNet model was used. The first two rows show the saliency maps where both model gave correct predictions. We can see from saliency maps that the Anisotropic model has more diffused saliency maps. The second two rows show the saliency maps where Anisotropic model gave correct predictions and ImageNet model gave wrong predictions. The failure of ImageNet model might be due to it not attending to whole object. The last two rows show the saliency maps where Anisotropic model gives incorrect predictions and ImageNet model gives correct predictions. Even in this failure mode, the Anisotropic model gives diffused saliency maps.

REFERENCES

- Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. 1, 2, 4, 5
- Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. 5

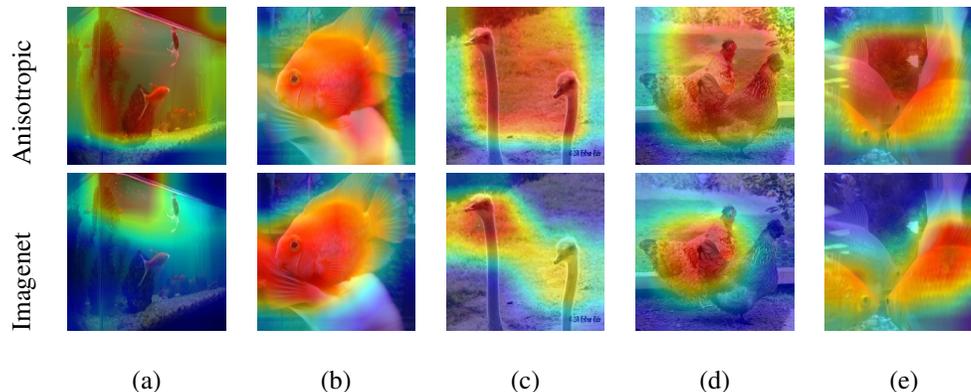


Figure 4: Saliency maps when our technique gives the correct prediction and baseline approach gives incorrect label. The top row gives the saliency maps for our model and the bottom one shows the corresponding saliency maps for the model trained on imagenet alone. We can see from saliency maps that Anisotropic model has bigger saliency maps which might be the reason for the correct prediction.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 3

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1422–1430, 2015. 2

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2009. 5

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*, abs/1811.12231, 2018. 2, 5

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 4, 5

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019. 4, 5

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60:84–90, 2017. 1

T. Nathan Mundhenk, Daniel Ho, and Barry Y. Chen. Improvements to context based self-supervised learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9339–9348, 2017. 1

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. *ArXiv*, abs/1603.09246, 2016. 1, 2, 4, 5

Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 4, 5

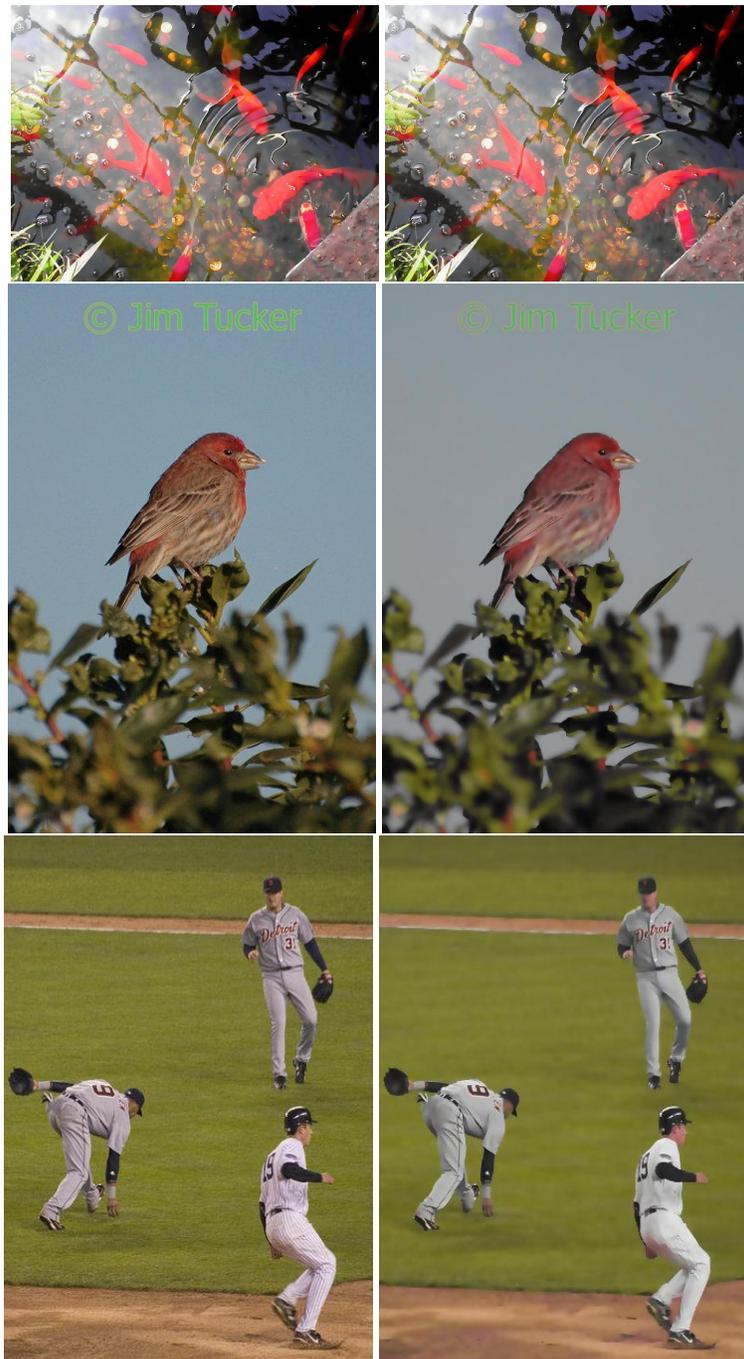


Figure 5: Original images (left) and images obtained after anisotropic diffusion (right). Most of the texture information in the images has been smoothed out by the filter while retaining the shape information. This forces the network to capture higher-level semantics without relying on low-level texture cues

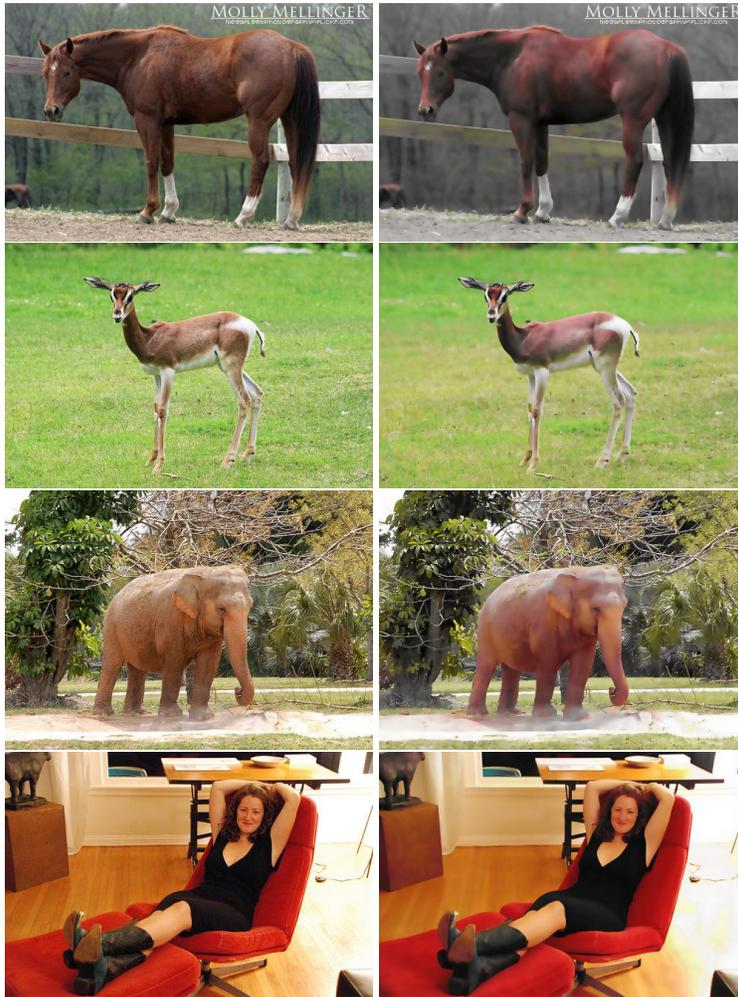


Figure 6: Original images (left) and images obtained after anisotropic diffusion (right). Most of the texture information in the images has been smoothed out by the filter while retaining the shape information. This forces the network to capture higher-level semantics without relying on low-level texture cues

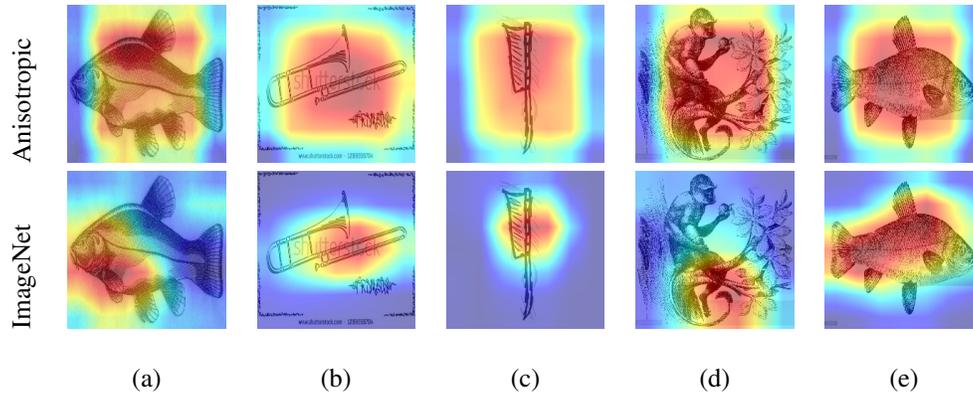


Figure 7: Saliency maps on few randomly selected images from Sketch-ImageNet. We can see from saliency maps that Anisotropic model has bigger saliency maps which might be the reason for the correct prediction.

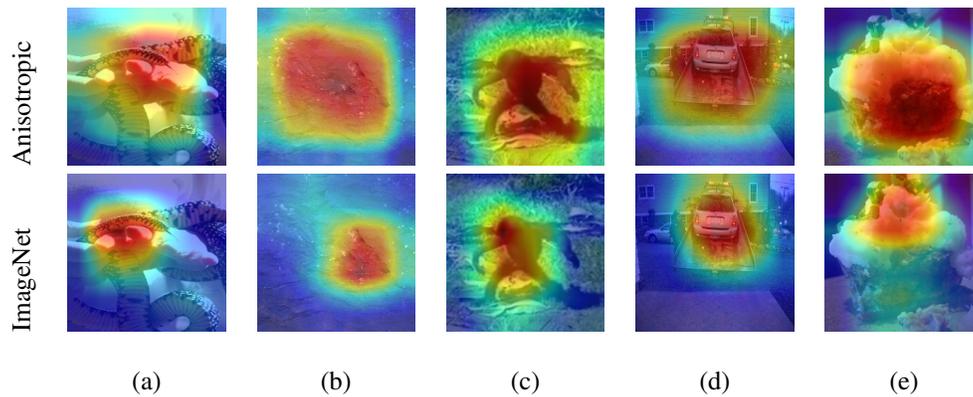


Figure 8: Saliency maps when Anisotropic Model had correct predictions and ImageNet model has wrong predictions.

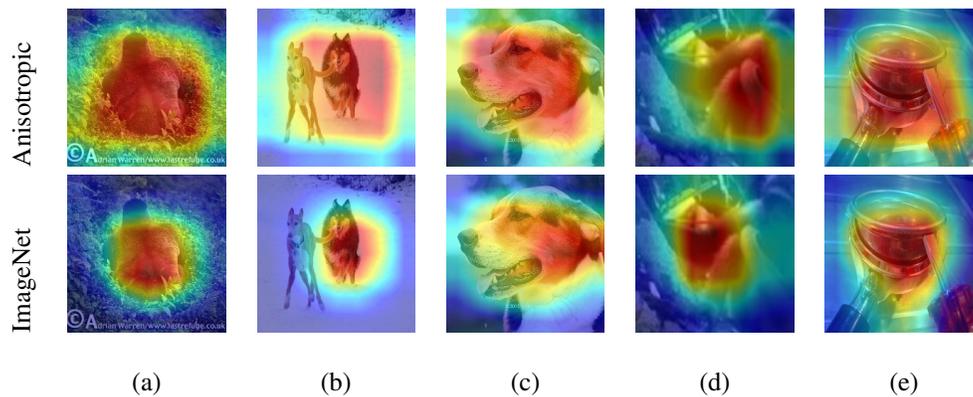


Figure 9: Saliency maps when both model have wrong predictions.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2016.

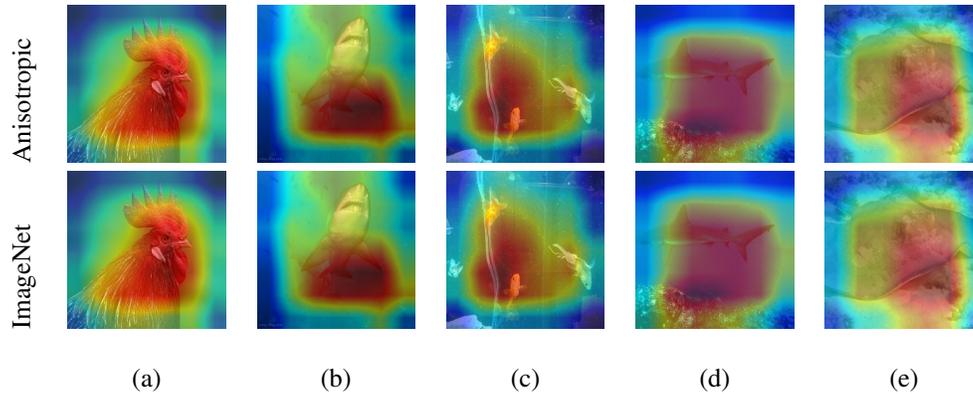


Figure 10: Saliency maps when both model have correct predictions.

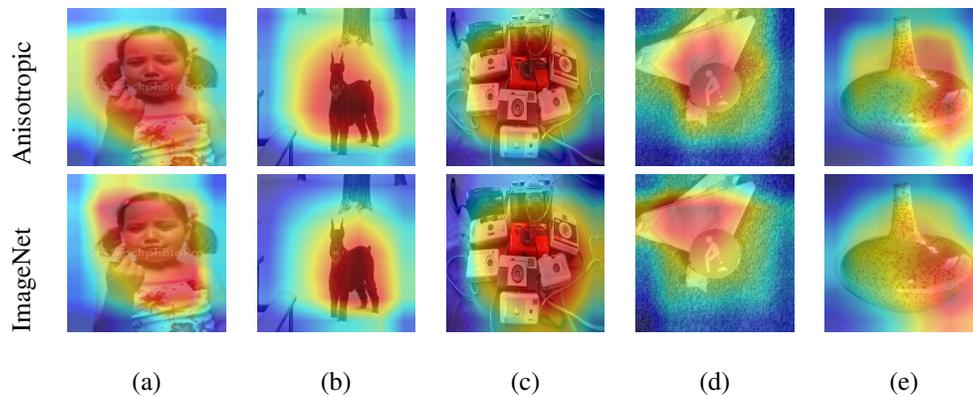


Figure 11: Saliency maps when ImageNet model has correct predictions and Anisotropic model has wrong predictions.

Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, Apr 2017. ISSN 2160-9292. doi: 10.1109/tpami.2016.2572683. URL <http://dx.doi.org/10.1109/TPAMI.2016.2572683>. 4

Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pp. 839–846, 1998. 3