Can self-supervised localizations improve out of distribution robustness?

Yenho Chen Kshitij Pisal Shaunak Halbe
Georgia Institute of Technology
{yenho, kpisal3, shalbe9} @gatech.edu

Abstract

Despite achieving superior performance on curated benchmarks on computer vision tasks, deep learning models are brittle when deployed into real-world settings where data-distributions are non-stationary. Poor generalization is often the result of large neural network models learning representations that overfit to specific datasets rather than capturing the underlying principles of a particular task. We hypothesize that incorporating self-supervised learning into a supervised task can encourage models to avoid these narrowly successful strategies in favor of ones that are more robust. Specifically, we use attention maps from self-supervised vision transformers that fixate on foreground objects to improve image classification accuracy on an out-of-distribution test set. We compare in-distribution and out-of-distribution (OOD) accuracy in a variety of settings and discuss possible sources of the seen effects. Furthermore, we provide an adaptive fine-tuning approach that is simple to implement, yet dramatically improves OOD classification accuracy (6%)

1. Introduction

Deep neural networks (DNN) have successfully achieved expert-level performance on benchmarks across many domains [4, 7, 12–15, 20]. Despite reaching state-of-the-art, deep learning models often solve problems with strategies that rely on *shortcuts* [6]. That is, rather than learning features that capture the underlying principles of a task, DNNs often follow unintended strategies that exploit the particularities of a given dataset to perform well on the train and test sets but fail to generalize to out-of-distribution (OOD) test sets. Instead of prioritizing narrow success, application areas that involve risky decision-making prioritize methods that avoid shortcut learning and can learn robust and generalizable representations instead.

Self-supervised learning encourages models to avoid shortcut strategies by constructing an internal representation using supervisory signals that leverage the structure of the data itself and do not require explicit labels. These techniques are very successful across a variety of domains. For example in NLP, SOTA models are trained through self-supervision, with techniques such as next-word prediction and masked language modeling [1, 4]. In neuroscience, self-supervised learning can construct latent representations of neural signals by associating adjacent time points to create meaningful data visualizations [18]. In computer vision, self-supervision is an effective tool for pretraining, and auxiliary tasks such as colorizing grayscale images can dramatically improve downstream tasks of interest such as object tracking [3, 24]. Importantly, self-supervised approaches improve robustness and uncertainty estimation without requiring larger models or more data. As a result, self-supervision as a form of implicit regularization against shortcut learning can be of interest to a broad range of applications including healthcare, self-driving cars, and weather prediction.

1.1. Project Goals

Our goal is to develop a strategy that mitigates shortcut learning by encouraging the model to form robust representations through self-supervision. We focus on improving image classification accuracy on OOD test sets without any data augmentation. Instead, we rely on additional information provided by self-supervised vision transformers (ViTs) which have been demonstrated to perform object localization without ever explicitly training for it. Specifically, we give a ViT an image and generate an object localization map which can be used to determine regions of importance. The masked images are used as inputs to a downstream classifier which will predict a categorical variable indicating the predicted class. We quantify the effects of self-supervision on robustness by comparing the test performance between the classifier and a baseline across a variety of evaluation metrics. We expect that our models will have similar in-distribution test performance as the baseline, but the inclusion of a self-supervised signal will allow us to dramatically outperform the baseline on OOD test sets. In doing so, **we hope to improve classifier robustness on OOD classification for visual recognition systems** which is a crucial step to developing trustworthy computer vision models that can

be used in real-world applications. If successful, we will have explored a modern form of deep learning regularization which is of particular interest to application areas that involve high-risk decisions.

2. Related Works

The following works are relevant to model robustness and self-supervision for image data. We highlight important results and elaborate on how we intend to utilize or improve on previous work.

2.1. Localization Emerges from Self-Supervision

ViTs have demonstrated performance competitive with Convolutional Neural Networks on vision tasks. Self-supervision can be incorporated using the self-distillation with no labels (DINO) framework [2]. The model includes a student and teacher network which receives two different random transformations of an input image. Outputs of the teacher network are centered while outputs of the student network are not transformed. The self-supervision objective minimizes the distance between the output features of the student and teacher network. Interestingly, the resulting features produce attention maps that highlight salient foreground objects and can be used as high-quality features for k-NN classifiers and even object discovery [25]. Our work utilizes the emergent localization property of ViTs to provide unsupervised localization masks.

2.2. Self-Supervision Improves Model Robustness

[8] demonstrates that incorporating a self-supervision task into supervised training can improve model robustness. While there was little effect on overall accuracy, the additional information provided by self-supervision improves performance in corrupted data classification and OOD detection. These promising results were shown on simple self-supervised tasks such as standard projected gradient descent, which includes an inner loop that creates adversarial training samples, and an auxiliary rotation task, where the model learns to predict the rotation angle of an image but implicitly forms a robust representation for object shapes while doing so. We extend this work by using more powerful self-supervision approaches that leverage modern advances in deep learning. Rather than relying on simple image rotations, we use the emergent localization properties from ViTs as a self-supervisied training signal.

2.3. Data Augmented Invariant Regularization

DAIR is a class of methods that approach model robustness by altering the optimization objective with additional regularization terms that promote learning representations invariant to certain transformations [9]. The DAIR regularizer is based on the idea that a fully invariant model will have the same loss score for a sample and its augmented version and minimizes the distance between these two. This approach is similar to consistency regularization penalizes the distance between the predictive distribution of two augmentations of a reference sample [22]. Our approach is orthogonal to this class of methods since we do not use explicit regularization. Instead we rely on the implicit regularization provided by self-supervision. In doing so, we hope to avoid the costly hyperparameter tuning that is required to select an appropriate regularization strength.

2.4. Alleviating Sprurious Feature Learning

Models trained with Empirical Risk Minimization (ERM) tend to do well on samples lying in the training distribution. When evaluated on out-of-distribution data, ERM seems to perform miserably. This suboptimal performance of ERM is largely attributed to the learning of spurious correlations or "shortcuts" and has been studied in great details in recent works [6, 16]. Some works have focused on annotating worst-groups in datasets and using these annotations as supervision to enforce robust training. However, assuming access to group annotations is impractical in the real world. Hence, most recent works have explored ways to identify and exploit spurious correlations. [11] leverage the shortcut learning tendency of ERM to detect worst-groups within the training data. Further, they perform robust optimization on these groups by upweighting the corresponding samples.

In parallel to this, [23, 27] explore a more specific form of spurious correlations namely the foreground-background confusion scheme. These works share the same motivation as our project. Masktune [23] leverages input attribution techniques like GradCam to identify regions in the image that the model deems important. Further, they mask these regions to encourage the model to explore new features and hence reduce the reliance on spurious features. We extend this idea in our project by leveraging self-supervised saliency maps to find objects of interest and masking out the backgrounds. The adaptive training scheme that we implement achieves strong results for both in-domain and out-of-domain benchmarks.

3. Method

Fig 1 shows that our approach consists of three phases. First, we establish performance baselines on in-domain and OOD test sets by providing classifiers with oracle localization maps. Next, we assume that oracle masks are not available and replace them with attention maps provided by self-supervised ViTs. Finally, we leverage these attention maps to learn robust features using a careful fine-tuning strategy.

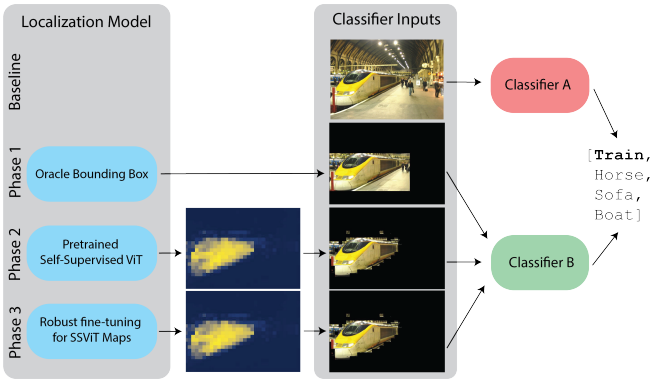

Figure 1. Our approach consists of three phases which we compare to a baseline. Classifier A receives raw images as inputs and includes no localization model (Top). In contrast, classifier B received oracle bounding boxes provided by the dataset in phase 1, attention maps from pre-trained self-supervised ViT in phase 2, and attention maps from fine-tuned self-supervised ViT maps in phase 3 for the same classification task. We hypothesize that the use of localization maps will improve OOD classification accuracy which implies that the model has learned a robust representation of the objects

Train	In-distribution Test	OOD Test
VOC	VOC	COCO
COCO	COCO	VOC

Table 1. Testing directions that we consider. In distribution test is defined as training and testing on data from a single dataset. OOD test is defined as training on one dataset and testing on the other. Thus we have two possible configurations for each metric as shown.

3.1. Establish Baselines using Oracle Maps

The goal of Phase 1 is to establish a baseline for the benefits of self-supervision. We want to demonstrate that classification performance improves when trained on images that have the background masked out. By removing extraneous information in the image, neural network models are encouraged to learn features that can distinguish between each of the object classes using information from only the objects themselves, rather than relying on other context information that can yield good performance but do not relate to the object classes. For the oracle masks, we use pre-annotated bounding boxes provided by popular object recognition datasets.

For a given image classification dataset, we split it into two portions: train and validation, where the training data consists of the datapoints that are used to fit neural network classifiers and the validation data is used to estimate in-distribution generalization accuracy. As a result, the label distribution of each split should be identical which can be achieved using a stratified splitting procedure for each label class. To estimate OOD generalization, we require a second dataset with the same label classes as the first, but is constructed using a different data collection procedure. The two different generating procedures for each dataset should result in quantifiable differences in their respective data distribution which will cause OOD accuracy to decrease when compared to the in-distribution accuracy.

For each dataset, we consider two types of classifiers: A which is trained only on the whole image, and B which is trained only on images with background masked out.

First, we determine the performance baseline for COCO to VOC generalization. Networks A and B are trained on the same images from COCO with the only difference being that the background is masked out using oracle bounding boxes in B. In-domain baseline is established by testing on the same held-out samples from COCO in both networks. Baseline OOD performance is evaluated by looking at the classification accuracy on held-out samples from the VOC dataset. The second set of experiments are identical to the first step, but establish a baseline in the other dataset direction - i.e. training on VOC and testing on COCO for in-domain and OOD baselines.

3.2. Extending Self-Supervised ViT

The use of pre-labeled oracle masks is unrealistic. In phase 2, our goal is to demonstrate that similar performance can be achieved using localization maps obtained from a self-supervised ViT pre-trained on a separate dataset, ImageNet. We use its emergent localization maps to replace the oracle masks from Phase 1 and repeat the same set of experiments, computing in-domain and OOD metrics in both dataset directions. In this phase, we will not fine-tune the ViT and use only use its output for the downstream classifier. These attention maps will be used to mask out uninformative regions in the inputs of network B. If successful, we will show that network A and B will have similar in-domain scores, but B will outperform A on OOD scores. The goal is to replace the oracle map with a deep learning solution that can achieve similar or better performance, but work in real-world settings where an oracle is not available.

3.3. Robust fine-tuning on SS-ViT Maps

In the final phase, motivated by the success of methods like Masktune [23] and Just-Train-Twice (JTT) [11] which employ a two-stage training scheme to enforce robust learning, we implement a smart fine-tuning scheme to boost our robust performance while preserving the performance on in-domain data. Our method is motivated by the efficacy of Self-Supervised methods like TokenCut in discovering the salient parts of an image. Models that focus on salient parts of images tend to do well on out-of-domain data due to lack of background correlations [27]. We leverage TokenCut to obtain images from the training data with back-

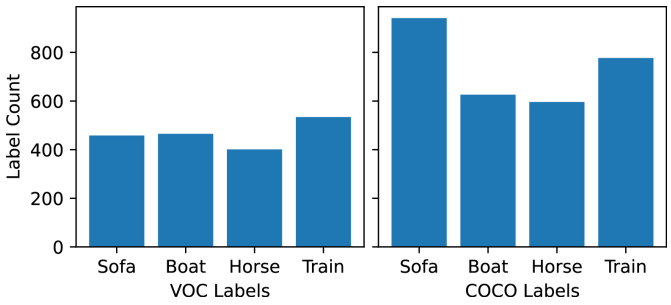

Figure 2. Label distribution for each dataset. VOC has a more balanced class distribution while COCO-mini's label distribution is faithful to the original COCO dataset statistics which includes a higher degree of label imbalance.

grounds masked out. Initially, we perform ERM training on the original images of a dataset (VOC/COCO). Next, we adapt the learning rate to a smaller value and perform one round of finetuning on the masked data obtained from either TokenCut or oracle annotations (See Table 2, Row 3 and 4).

4. Data

COCO Mini Dataset (Common Objects in Context) COCO [10] was constructed by Microsoft with the goal of advancing object recognition for objects placed in a broader context of a particular scene. In total there are 328k images with 91 object types and 2.5 million label instances. Images are collected and annotated using crowd workers and represent complex scenes that contain many common objects in their natural context. As such, a single scene contains potentially several object classes and instances of each class where each class instance may overlap with another. Bounding boxes for each instance are provided as part of the annotations.

Rather than using the entirety of the COCO dataset, we choose to use a miniaturized version of COCO for fast method development and quick iterations for our project. Randomly sampling images from COCO does not work as the dataset is a large collection of 80 classes with images having different object instances, image sizes and object sizes in images. To ensure that the reduced set is similar to the original set we seek to mirror the following characteristics: 1) Number of object instances per class in the sample, 2) Ratio of large, small and medium objects in the sample, and 3) Ratio of large, small and medium objects per class. Thus we use COCO-mini [17] which is a curated subset of the COCO 2017 train set aimed at reducing the computational costs of running experiments. COCO-mini is constructed so that it provides a recommended train and validation set that yields validation performance strongly correlated with the original dataset. Furthermore, object instance statistics are preserved in the reduced dataset to maintain the characteristics of interest listed above. In total, 25,000 images ($\approx 20\%$) are sampled to form COCO-mini which helps reduce the overall disk space and memory requirements.

PASCAL VOC Dataset Visual Object Classes (VOC) seeks to provide a standard dataset of images for an annual computer vision competition which was hosted every year between 2007 through 2014. This event is hosted by PASCAL (pattern analysis, statistical modelling and computational learning) which is in the EU Network of Excellence. The challenge consists of object classification and detection of instances. Images are collected from flickr, a photo-sharing website, and are selected without the challenge in mind to get a better representation of images in their natural settings. In total there are 11,540 images containing 20 object classes with potentially multiple categories and instances in each image. Furthermore, bounding boxes for each category are provided using the TKK method as specified in [5]. Since VOC is a much smaller dataset compared to COCO, we decide to consider the full dataset when developing our pipeline and do not reduce its size.

Data Preprocessing For both datasets, we restrict the number of classes to a subset of the overall classes. There are two selection criteria. First, selected classes had to exist in both datasets so that we can accurately test OOD accuracy. Second, we choose the classes that are least likely to be overlapping in a single image so the effect of overlapping objects would not be a confound in our experiments. The four classes selected were: ['boat', 'sofa', 'horse', 'trains']. After reducing, we retain 2368 images in COCO and 1920 images in VOC. We also quantify the distributional shift between the two datasets by looking at changes in class label and pixel distributions. In Figure 2 we show that VOC contains approximately even number of counts between all class labels. In contrast, COCO-mini has much greater label imbalance which reflects the requirement of matching the original COCO dataset properties. In Figure 3, we show that although the total pixel counts for each channel have the same shapes for each category, the numerical values are suppressed in COCO which results in flatter looking images. This suggests that there is much less contrast in COCO images for each respective category which may be the result of selecting for naturalistic scenes compared to scraping a photo-sharing website that can contain a wider variety of lighting conditions.

5. Experiments and Results

5.1. Baseline Analysis

We split VOC and COCO into training and testing sets with a 80:20 ratio. We based our model architecture on VGG16 [21] and use only 2 VGG blocks per network. Each block consists of 2 convolutional layers to extract 2D im-

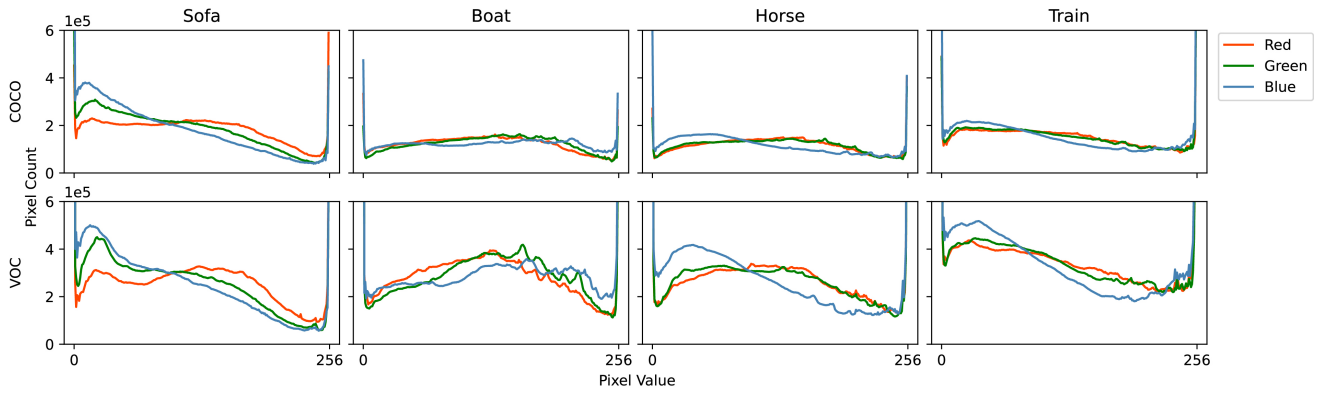

Figure 3. Pixel distribution of each category for each dataset. For each class, pixel distribution have similar shapes between datasets. However, COCO images are shown to have less contrast as a result of their data collection procedure which prioritizes naturalistic scenes. VOC images are collected from flickr and allows for more variety of scenes which is shown in its higher contrast.

Figure 4. Confusion matrices for Vanilla classifier, Network B trained on images masked with oracle maps (phase 1) and localizations provided by pretrained self-supervised ViT (phase 2)

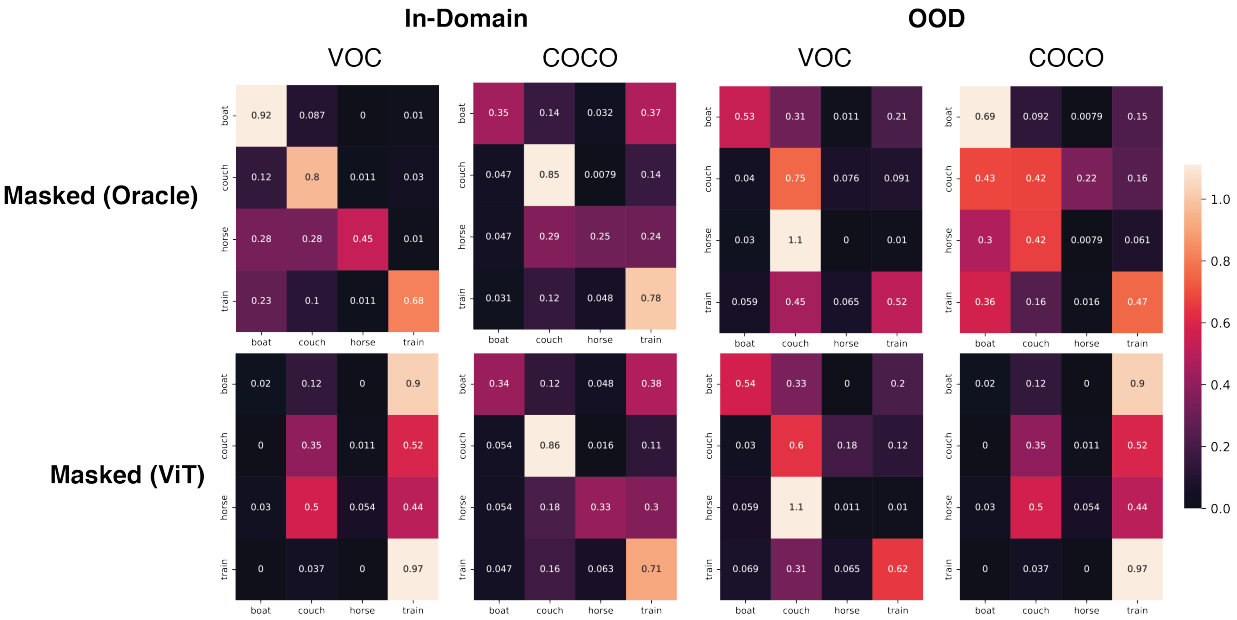

Figure 5. Confusion Matrices for Adaptive Training scheme. First row represents the results obtained by using oracle masks while fine-tuning and second row represents for ViT generated masks.

age features, followed by a max-pooling layer to reduce the dimensionality of the features, then a batch normalization layer to stabilize learning, and a ReLU activation function to introduce nonlinearity in the model. We chose a reduced VGG architecture as a form of implicit regularization against overfitting due to our relatively small dataset size. We also consider models with 1 and 3 VGG blocks and found that one block does not have enough depth to adequately separate the classes while 3 VGG blocks yields poor in-domain accuracy as a result of overfitting. Thus 2 VGG blocks was the best choice for this project. All networks are trained for 30 epochs on their corresponding training set (either COCO or VOC). We consider the performance of the network on the in-domain test set to ensure that our model is not overfitting and the OOD test set as an estimate for empirical risk. In the VOC $\rightarrow$ COCO setup, we train Network A on the original images and Network B on masked images from the VOC train set. We evaluate these networks for in-domain performance on the corresponding test splits of VOC. To measure OOD performance, we evaluate these networks on the test split from the COCO dataset. Symmetrically, we repeat this procedure for the other dataset direction: COCO$\rightarrow$ VOC.

Table 2 shows these results. Intuitively, we expect Network B to generalize better than Network A owing to the masking of background signals which are believed to be spurious. We observe that Network B outperforms Network A on in-distribution data in VOC, but not COCO datasets. However, the OOD results show improvement for both VOC and COCO. Looking at the confusion matrices, we see good performance for Network A since most predictions lie on the main diagonal indicating that the model is mostly correct in its predictions in both datasets. Similarly, network B show superior in-domain predictions with a strong diagonal of correct predictions. However in the OOD sets, network A has unexpected behavior, being heavily biased to predict "couch" in the COCO $\rightarrow$ VOC direction and "train" in the other direction. A similar effect is seen in network B is heavily favoring "train" and "boat" predictions in the VOC $\rightarrow$ COCO direction and "couch" in the other direction. One possible explanation is that the network learns the prior distributions for the class labels of their respective datasets and incorrectly applies them on the OOD data. Figure 2 shows that a network trained on COCO will have a much higher probability of seeing an image with a "couch" compared to other classes while a network trained on VOC will have a higher probability of seeing an image in the "train" class.

5.2. Attention maps from Self-supervised ViT

In phase 2, we replace the oracle maps with attention maps from self-supervised ViT. Specifically, we use the TokenCut method [26] which takes attention maps produced by pretrained vision transformers and produces a segmentation map that distinguished between foreground and background. Images are segmented using spectral clustering with generalized eigendecomposition which reformulates the task as a graph-cutting problem. The analysis in this phase also follows the ones done in phase 1, where we compare the in-domain and OOD accuracy between network A

		In-domain Accuracy		OOD Accuracy	
Phase	Strategy	VOC	COCO	VOC	COCO
-	A	68%	67%	35%	33%
1	B	**70%**	54%	**38%**	**35%**
2	B	37%	48%	35%	**36%**
3	B (ViT Maps)	36%	59%	**44%**	**36%**
3	B (Oracle Maps)	**71%**	61%	**44%**	**40%**

Table 2. Results. In-domain accuracy measures the performance on test-split of the dataset that was used for training. Out-of-domain accuracy is measured for the test split of the dataset which is not used for training. Bolded represents superior accuracy scores against Network A. We see a tradeoff between in-domain and OOD accuracy as a result of our implicit regularization.

and B.

Table 2 shows that the in-domain performance for network B decrease in comparison to network A and the result in phase 1. This is expected since oracle maps represent the upper bound on performance benefits since localization maps are hand-curated by crowd-sourced workers. In contrast, localization maps provided by ViT's are noisy and often fixate on areas of the image that do not well capture the class label of interest. For example, if a human and a couch are in the same image, ViT's will treat the human as the foreground and the couch as the background, thus eliminating information that can be used to predict the true class label. Surprisingly, OOD accuracy does stays the same in VOC and even improves in the COCO dataset. This suggests that noisy maps can provide enough regularization to encourage more robust representations.

5.3. Careful Finetuning on SS-ViT maps improves Robustness

We observe a strong boost in performance (**6%**, See 2) on out-of-domain data which is the focus of our paper and a modest boost on in-domain data. Intuitively, our two-staged approach preserves the in-domain discriminative power of the model and improves robust performance by finetuning for a single epoch on masked data. The smaller learning rate helps in not diverging from the in-domain solution. Particularly, we decrease the learning rate to 1/10th of the final learning rate from the ERM method. We corroborate the observations of [23] that fine-tuning for longer epochs or using higher learning rate diverges from the optimum. As compared to masktune, our approach does not require gradient attribution techniques like GradCam [19] to determine saliency maps and instead work with self-supervised maps that can be precomputed thereby increasing the training efficiency of the model in contrast to the former approach. Secondly, if the self-supervised method is jointly trained with the classifier, we see our method achieving even stronger results.

6. Conclusion

We demonstrate that self-supervised localization maps can improve the OOD accuracy of a neural network classifier. In phase 1, we establish a baseline using oracle maps and show improved in-domain and unaffected OOD accuracy. In phase 2, we use attention maps provided by pretrained self-supervised vision transformers to mask out background objects and improve classifier accuracy on OOD test set. We show that as OOD accuracy increases, in-domain accuracy decreases; suggesting that there is a tradeoff between the two. This makes intuitive sense. Our implicit regularization approach prevents overfitting to the specific data distribution of the train dataset which results in lower in-domain accuracy, but yields better performance on OOD samples. In Phase 3, we show that executing a careful finetuning strategy on attentions maps obtained from SS-ViTs provides a strong boost in OOD performance while maintaining the in-domain accuracy. Future work can consider implementing different forms of joint training that scales our initially promising results to problems of larger size (> 4 classification classes) or leverage self-supervised implicit regularization for different data types such as time-contrastive learning for modeling time-series.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 1

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 4

[6] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, nov 2020. 1, 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 1

[8] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty, 2019. 2

[9] Tianjian Huang, Shaunak Halbe, Chinnadhurai Sankar, Pooyan Amini, Satwik Kottur, Alborz Geramifard, Meisam Razaviyayn, and Ahmad Beirami. Dair: Data augmented invariant regularization, 2021. 2

[10] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[11] Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information, 2021. 2, 3

[12] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017. 1

[13] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1

[14] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019. 1

[15] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022. 1

[16] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2019. 2

[17] Nermin Samet, Samet Hicsonmez, and Emre Akbas. Houghnet: Integrating near and long-range evidence for bottom-up object detection. In *European Conference on Computer Vision (ECCV)*, 2020. 4

[18] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioral and neural analysis. *arXiv preprint arXiv:2204.00673*, 2022. 1

[19] 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. *International Journal of Computer Vision*, 128(2):336–359, oct 2019. 7

[20] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016. 1

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 4

[22] Jihoon Tack, Sihyun Yu, Jongheon Jeong, Minseon Kim, Sung Ju Hwang, and Jinwoo Shin. Consistency regularization for adversarial robustness, 2021. 2

[23] Saeid Asgari Taghanaki, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi-Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore, 2022. 2, 3, 7

[24] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018. 1

[25] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 2

[26] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, June 2022. 6

[27] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition, 2020. 2, 3