Masked Images Are Counterfactual Samples for Robust Fine-tuning

Yao Xiao Ziyi Tang Pengxu Wei * Cong Liu Liang Lin
Sun Yat-sen University
{xiaoy99,tangzy27}@mail2.sysu.edu.cn {weipx3, liucong3}@mail.sysu.edu.cn
linliang@ieee.org

Abstract

Deep learning models are challenged by the distribution shift between the training data and test data. Recently, the large models pre-trained on diverse data demonstrate unprecedented robustness to various distribution shifts. However, fine-tuning on these models can lead to a trade-off between in-distribution (ID) performance and out-of-distribution (OOD) robustness. Existing methods for tackling this trade-off do not explicitly address the OOD robustness problem. In this paper, based on causal analysis on the aforementioned problems, we propose a novel fine-tuning method, which use masked images as counterfactual samples that help improving the robustness of the fine-tuning model. Specifically, we mask either the semantics-related or semantics-unrelated patches of the images based on class activation map to break the spurious correlation, and refill the masked patches with patches from other images. The resulting counterfactual samples are used in feature-based distillation with the pre-trained model. Extensive experiments verify that regularizing the fine-tuning with the proposed masked images can achieve a better trade-off between ID and OOD performance, surpassing previous methods on the OOD performance. Our code will be publicly available.

1. Introduction

Deep learning has achieved impressive advances in various tasks on computer vision. Despite the remarkable performance on benchmark datasets, the deep models are challenged by the distribution shift between the training data and test data [5, 15, 16, 35]. It is commonly assumed that the training and test samples follow the same distribution, which may not hold in real-world applications due to the unpredictable change of lighting condition, viewpoints, backgrounds, *etc*. A series of works attempted to improve the robustness of deep models to the distribution shift (or OOD robustness), but it is still rather under-explored [30, 40].

*Corresponding author.

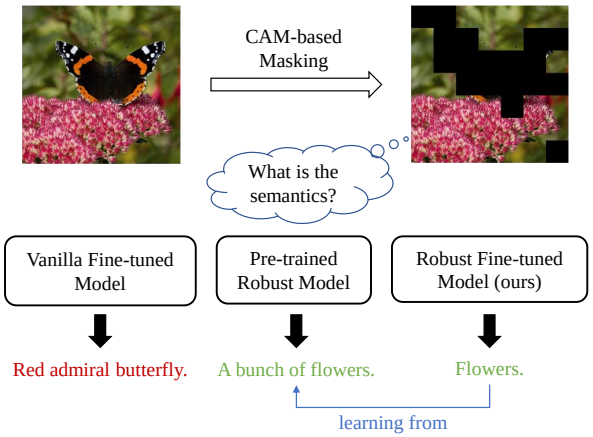

Figure 1. Illustration of our work. Vanilla fine-tuned models tend to learn spurious correlations that degrades the OOD robustness. To tackle this issue, our model learns from the pre-trained model on the counterfactual CAM-based masked images.

Recently, the large models pre-trained on diverse data demonstrate unprecedented robustness to various distribution shifts [8, 19, 34]. Hence, fine-tuning such pre-trained models on downstream tasks can be a promising approach to building robust models for different applications. However, it is found that while fine-tuning improves the performance on in-distribution (ID) data, it may reduce that on out-of-distribution (OOD) data [23,46]. To tackle this trade-off between ID and OOD data, several methods [23, 45, 46] have been proposed to improve both ID and OOD performance in fine-tuning. However, they do not explicitly address the OOD robustness problem, but implicitly preserve the robustness of the pre-trained model by constraining the distortion of pre-trained weights or using model ensembles.

In this paper, we revisit the issue of robustness degradation in fine-tuning from a causal perspective. We notice that a large-scale pre-trained model somewhat shows properties in causality and stays robust to OOD samples [44]. However, when fine-tuning on downstream tasks, a majority of the parameters of the model tends to be adjusted for the downstream task in fine-tuning due to the highly entangled

representation of images, arguably destructive to the generalizable knowledge [20, 37]. In contrast, distribution shifts are usually sparse in the underlying causal factorization of data generation process [7, 37]. In this low-dimensional case, if we know which variables vary with different data distributions in this factorization (*i.e.*, the non-stationary factors), we can achieve the OOD robustness by simply excluding their influence on final predictions of the model.

Specifically, we consider a Structural Causal Model (SCM) [33] for the object-centric image generation process, as depicted in Fig. 2. In this SCM, images are generated according to a non-stationary domain-relevant factor and a stationary semantic factor. Between them is a spurious correlation caused by a hidden non-stationary confounder that influences how the domain-relevant factor changes with the semantic one. To retain the OOD robustness, a fine-tuning model should avoid mapping non-stationary domain-relevant features to the predicted semantics.

To this end, we propose to fine-tune the models with masked images, which serve as counterfactual samples breaking the spurious correlation. Training on these samples helps preserve the stationary and generalizable knowledge of the pre-trained model. Concretely, we either mask the patches that contribute most to the label (*i.e.*, the main object), or mask those with the least contribution (*e.g.*, the context), which can be implemented based on class activation map (CAM) [9, 49]. Such image masking forms an manipulation of a factual image and produces a counterfactual sample. Since the pre-trained model can better disentangle invariant features across domains, we require the fine-tuning model to learn from the pre-trained model on these counterfactual samples, as illustrated in Fig. 1. Furthermore, we argue that simply dropping the masked patches may be insufficient to alleviate the risk of fitting spurious correlations, and we propose to refill the masked patches with those from other images.

We study different combinations of masking strategies (*e.g.*, masking the object or the context) and refilling strategies (*e.g.*, filling with patches from single or multiple images). Experimental results suggest that most of the strategies are applicable to the construction of counterfactual samples that help improve the robustness in fine-tuning, while masking the object generally achieves the best robustness. Compared with existing methods [23, 45, 46] on fine-tuning CLIP [34] models, our approach achieves better average accuracy on various OOD datasets without relying on model ensembles or weight constraints. We also find that taking the weight-space ensemble of the zero-shot model and our fine-tuned model following WiSE-FT [46] hardly improve the trade-off between ID and OOD accuracy, which contradicts with previous observations and implies that our approach may produce essentially different models as compared with vanilla fine-tuning.

2. Related Works

2.1. Causal Perspective on OOD Robustness

Causality holds the promise of OOD robustness as it has the property of robustness under interventions that underlie distribution shifts [2, 37]. To learn the latent causal mechanism in visual data and attain OOD robustness, a branch of methods [21, 29, 38] target at learning domain-invariant causal representations and seek guarantees on generalization. Some approaches approximates this goal based on invariant risk minimization [1, 3], risk extrapolation [22], adaptation speed [7], or variational Bayes [26]. Ilse *et al.* [18] help models escape from spurious correlations by training on handcrafted analog interventional data. In this work, we aim to explore how to preserve the OOD robustness of a pre-trained model equipped with generalizable knowledge during transfer, based on the analysis of an underlying data generation causal mechanism.

2.2. ID-OOD Trade-off in Fine-tuning

While a series of large pre-trained models exhibit strong OOD robustness [8, 19, 34], empirical results suggest that fine-tuning these models on downstream data may degrade the robustness [23, 34, 46]. It is also theoretically justified that compared with linear probing, vanilla fine-tuning may increase the ID performance while decrease the OOD performance due to feature distortion [23]. To tackle this ID-OOD trade-off, LP-FT [23] applies linear probing before fine-tuning; Calibrated Ensemble [24] and WiSE-FT [46] propose to take the ensemble of the pre-trained (zero-shot) model and vanilla fine-tuned model in output-space and weight-space, respectively. Model Soup [45] also improves the trade-off via weight-space ensembles, but it utilize multiple models fine-tuned with different hyper-parameters. While these methods are empirically effective in improving both ID and OOD performance, they only implicitly preserve the robustness of pre-trained model by constraining the deviation of the downstream model from the pre-trained one. Instead, our approach is based on explicit causal modeling on OOD robustness problem and the use of counterfactual samples in fine-tuning.

2.3. Learning with Masked Images

Masked image modeling [4, 14, 43, 47] has been proved to be an effective approach to vision model pre-training, where random sampling is a common strategy for masking. In our task, we find that learning with CAM-based masking can better address the spurious correlation in images. Similar to our approach, CSS [10] synthesizes counterfactual images for visual question answering (VQA) by masking out critical objects, for which the corresponding answer is changed to its negative; SwapMix [13] swaps the the context objects in the feature space to reduce the reliance of VQA models

on visual context. In this paper, we apply masking on images and refill the masked regions with content from other images, which is inspired by CutMix [48]. Besides, instead of manually constructing the labels for masked images like CSS or CutMix, we take the feature representations of the pre-trained model on the masked images as the supervision for the fine-tuning model.

3. Method

3.1. Revisit OOD Robustness of Fine-tuned Models from a Causal Perspective

When fine-tuning a pre-trained model on downstream tasks in real-world applications, it is usually insufficient to generalize only to in-distribution (ID) data. There are also increasing demands for robustness to distribution shifts or generalization to out-of-distribution (OOD) data. However, for a pre-trained large model that has been trained on data from diverse distributions and obtained generalizable knowledge (*e.g.*, CLIP [34]), fine-tuning can lead to dramatic degradation of robustness [23].

This problem can be interpreted from a causal perspective. Suppose that we correctly factorize the data generation process in which variables are causally connected. Then, the discrepancy between ID and OOD (or sub-population) data generation process is usually sparse in this causal factorization [37, 38], *e.g.*, few variables' priors change. The joint distribution that we can observe thereupon shifts. However, without modeling the causal hierarchy in data, the pre-trained models naturally learn dense and entangled features to represent the concepts in the causal factorization. Although distribution shifts are sparse in causal factorization, these models have to densely adapt their parameters in transfer to the downstream data [20]. In this case, damage to generalizable knowledge is inevitable, which explains the sharp decrease of robustness to distribution shifts.

The core challenge in modeling the underlying causal factorization of visual data is the dense changes between ID and OOD data, despite the sparsity in causal factorization. For better robustness, a possible solution is to reason backward which visual features are unchanged under distribution shifts, also known as *stationary features* [37]. Without OOD test data in hand to identify stationary features, we seek to model the most important stationary features as the semantic features in the downstream object-centric task, and reduce non-stationary features to domain-relevant features. Finally, we establish our causal model as in Fig. 2.

3.1.1 The SCM for Image Generation Across Domains

To illustrate the proposed method, we first introduce in Fig. 2 a Structural Causal Model (SCM) [33] to model the underlying mechanism of object-centric image data generation across diverse domains, which is agreed with previous works [18, 39] despite some slight modifications.

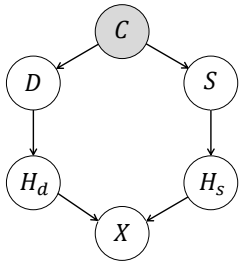

Figure 2. The causal graph of underlying object-centric image generation process across domains. C: confounder; D: domain; S: object semantics; H_d: (non-semantic) domain representation; H_s: semantic representation; X: image.

Formally, an SCM describes a directed acyclic graph composed of a set of endogenous variables V and a set of exogenous variables U. An endogenous variable is the variable whose value is determined by other variables. Exogenous variables correspond to unobserved influences, usually considered independent of each other. In Fig. 2, a set of endogenous variables are involved in the SCM, defined as $V = \{D, S, H_d, H_s, X\}$. There is also an exogenous variable for each endogenous variable, *e.g.*, U_{H_s} that possibly corresponds to the object pose and influences the generation of H_s. Here, we do not show them for simplicity, but only consider an important exogenous variable, the confounder C. We denote the set of exogenous variables by $U = \{C, U_D, U_S, U_{H_s}, U_{H_d}, U_X\}$.

In this SCM instance, C denotes an unobservable confounder variable, which is an exogenous variable shown in gray. For instance, it can be some specific time or space. D is a domain variable containing information varied with the domain. S is a semantic variable, *e.g.*, the category of the object. In our setting, D and S are observable endogenous variables shown in white. Correspondingly, H_d is the domain representation and H_s is the semantic representation, *i.e.*, they indicate how D and S are displayed in the image space, respectively. Finally, X is the image generated by the interaction of H_d and H_s.

3.1.2 Spurious Correlation and OOD Robustness

As depicted in Fig. 2, There exists a backdoor path between H_d and H_s, *i.e.* $H_d \leftarrow D \leftarrow C \rightarrow S \rightarrow H_s$, making them spuriously correlated. Collecting data for a downstream task from a single source or environment may lead to strong spurious correlations between the semantic part H_s and domain-relevant part H_d. In other words, there can be strong selection bias [29]. For example, in a downstream insect classification dataset, the *red admiral butterflies* may all be captured sitting on flowers, as shown in Fig. 1. However, this spurious correlation does not necessarily hold in

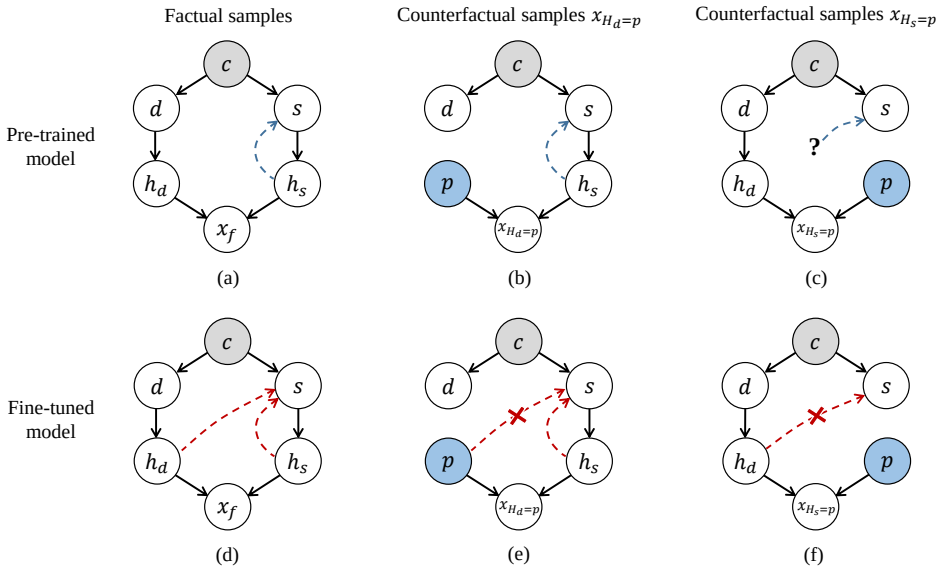

Figure 3. Conceptual comparison of pre-trained and fine-tuned models on what they depend on for predicting the semantic label of different samples. The robust pre-trained model is expected to depend only on h_s, while the fine-tuned model may associate both h_d and h_s to the semantics s. This may result in different predictions on counterfactual samples for the two models.

OOD data [6]. Moreover, in reality, what makes the label *red admiral butterfly* is the butterfly itself in the image space (H_s) rather than the flowers involved in H_d. Therefore, the model is required to depend only on H_s for the prediction of S to attain OOD robustness.

For the reasons we have discussed earlier, even if a pre-trained large model has been endowed with reasonable knowledge to distinguish H_s and H_d (Fig. 3 (a)), the model is still susceptible to the spurious correlations during fine-tuning (Fig. 3 (d)). As a result, the transfer dramatically degenerates the pre-trained model's OOD robustness.

From this perspective, a possible way to tackle the issue of robustness degradation is to break the spurious correlations in downstream data via specific manipulation, and explicitly requires the fine-tuning model to distinguish H_s and H_d following the pre-trained model. In the following sections, we propose a masking-based image manipulation method and indicate why it can be used to hinder the fine-tuning model from depending on H_d.

3.2. Masked Images as Counterfactual Samples

To break the spurious correlation in training images, we propose to mask out or replace certain regions of the images so that H_d or H_s is (partially) manipulated, which results in counterfactual samples. Formally, given an observational sample x and the SCM M, if we assign d' to variable D, the resultant counterfactual sample can be denoted as:

$$x_{D=d'}(u) = x_{M_{D=d'}}(u) \tag{1}$$

where $M_{D=d'}$ is the SCM instance in which d' is assigned to D, and u denotes the values of the exogenous variables.

We illustrate how the pre-trained and fine-tuned models may process two kinds of counterfactual samples differently in Fig. 3. First, for counterfactual samples whose domain-relevant representations h_d are replaced by other context p (*i.e.*, $x_{H_d=p}$ in Fig. 3 (b,e)), the fine-tuned model tends to attribute its prediction on both p and h_s since it has learned the spurious correlations. Hence, its prediction can be misled by p. Differently, the robust pre-trained model which can distinguish the semantic and non-semantic factors is not affected by p. Second, for counterfactual samples whose semantic representations h_s are replaced by p (*i.e.*, $x_{H_d=p}$ in Fig. 3 (c,f)), the fine-tuned model can still predict the original semantics s from h_d, while the pre-trained model cannot due to the missing semantic representations h_s. In both cases, counterfactual samples may lead to different predictions between the pre-trained and fine-tuned model.

Now the question is how to leverage these counterfactual samples to preserve the OOD robustness in fine-tuning. Directly training on these counterfactual samples with the labels of corresponding factual samples is unsuitable, since the original semantic information may be distorted. Given that the pre-trained models can have substantial power to capture semantic cues, their image-level feature representations usually contain rich semantic information. We thereby seek an alternative solution in which these samples are used in distillation. This way, the fine-tuning model learns to mimic the pre-trained model in feature representation. Formally, we denote the image encoders of the pre-trained and fine-tuning model by $\hat{f}$ and f, respectively, and denote the classification head of the fine-tuning model by g. Then, the

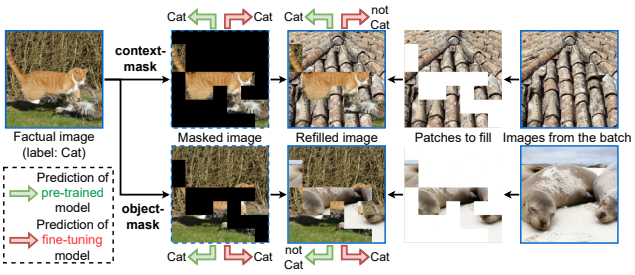

Figure 4. Illustration of the mechanism of masking and refilling.

overall training objective can be written as follows:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}\big(g(f(x)), y\big) + \beta \mathcal{L}_{\text{MSE}}\big(\hat{f}(x_{cf}), f(x_{cf})\big), \quad (2)$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, $\mathcal{L}_{\text{MSE}}$ is the mean squared error, x is the raw image, y is the label for x, x_{cf} is the counterfactual image, and β is a constant factor.

In the remaining parts of this section, we will discuss the proper design of masked images and show that it is nontrivial to construct the counterfactual samples x_{cf} that can be used to effectively improve the OOD robustness.

3.2.1 Choosing the Masked Regions

To generate counterfactual samples, we need to select the masked regions in an image where the contents can be manipulated. Recent large-scale pre-trained models [19, 34] usually adopt a Vision Transformer-based architecture [12], dividing input images into a series of patches. Hence, we consider patch-based masking in this paper. A basic strategy is to randomly sample a proportion of patches (*random-mask*), which is adopted by recent pre-training methods like MAE [14]. However, random sampling is content-agnostic. Although both H_s and H_d may be manipulated in this scheme, it tends to leave some spurious correlations in the unmasked regions. It is found that an MAE model can achieve semantically plausible reconstruction from the randomly masked images with a high masking rate (*e.g.*, 75%), suggesting that such random masking may preserve a significant amount of the information of the whole images. Therefore, randomly masked images may be less suitable for improving OOD robustness.

Alternatively, we investigate two content-based strategies that are complementary: (1) *context-mask*: masking the patches that are least relevant to the label (*i.e.*, H_d, usually the context); (2) *object-mask*: masking the patches that are most relevant to the label (*i.e.*, H_s, usually the main object). To measure the contribution of each image patch to the label, we generate the class activation map (CAM) [9, 49] based on the fine-tuning model. Then, image patches are separated into two groups by a constant threshold t. In such a scheme, the semantic-relevant regions (where their activation values are larger than the threshold t) can coarsely serve as H_s, and non-semantic regions as H_d.

3.2.2 Refilling the Masked Images

After choosing the patches to mask, simply dropping them following MAE [14] can lead to a counterfactual image, *i.e.*, $x_{H_d=\emptyset}(u)$. However, we argue that this strategy (abbreviated as *no-fill*) can be insufficient for the construction of effective counterfactual samples for improving the robustness, and refilling the masked regions can tackle this issue.

Concretely, an effective counterfactual sample should cause contradictions between the pre-trained and fine-tuning model, so that the latter is regularized by the distillation loss in Eq. (2) to depend less on non-semantic parts of images for the prediction of semantics. However, masking without refilling may not construct such samples, as depicted in Fig. 4. Specifically, *context-mask* may not produce contradictions since both models can predict semantics from the object, so we need refilling to bring some conflicting context that disturbs the prediction of the fine-tuning model. *Object-mask* alone can cause contradictions in theory, as the fine-tuning model could still predict the original semantics from the context, while the pre-trained model could not. However, due to the imperfect masking in practice, the pre-trained model may still recognize the object from its unmasked parts. Then, refilling can further distort the original semantics to ensure contradictions. Hence, refilling can help to construct more effective counterfactual samples for the proposed fine-tuning approach.

We consider two basic strategies to select the patches to fill: (1) *single-fill*: select the patches in the corresponding positions of a single image randomly sampled from the training batch; (2) *multi-fill*: for each patch to fill, independently sample a source image from the batch and select the patch in the corresponding position. All combinations of masking and refilling strategies are illustrated in Fig. 5.

4. Experiments

4.1. Setup

Datasets. We focus on fine-tuning the pre-trained model on ImageNet [36], and evaluate the OOD robustness of the models on five datasets: ImageNet-V2 [35] is a new test set for ImageNet collected following the original protocol. ImageNet-R [15] consists of various renditions (*e.g.*, art, cartoons) of 200 ImageNet classes; ImageNet-Sketch [42] contains sketches for each of the 1000 ImageNet classes; ObjectNet [5] is a test set with 113 overlapping classes with ImageNet, which is collected to show objects from new viewpoints on new backgrounds; ImageNet-A [16] contains natural images that are misclassified by models trained on ImageNet, covering 200 ImageNet classes.

Evaluation. Following [46], we use the top-1 accuracy as the metric of performance on ID and OOD data. For the five OOD datasets, we report the top-1 accuracy on each dataset, as well as the average OOD accuracy computed by

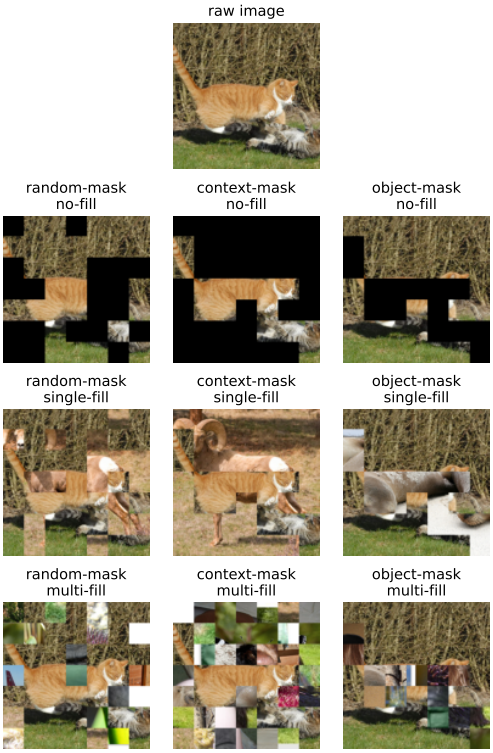

Figure 5. Illustration of different masking and refilling strategies.

averaging the accuracy on the five datasets. For datasets covering a subset of the ImageNet classes, the top-1 prediction of a model is taken as the class with highest probability among the class subset (instead of all ImageNet classes).

Implementation details. Unless otherwise specified, we use the ViT-B/32 model [12] pre-trained via CLIP [34]. The classification head of the zero-shot model is constructed from the pre-defined text prompts used by CLIP, and we adopt the implementation given by [46]. To obtain accurate CAM scores for image masking, we apply the method proposed in [9], which is designed for attention-based models, including ViT. For fine-tuning on ImageNet, we mainly follow the routine of WiSE-FT [46]. Specifically, we use the AdamW optimizer [28] with a batch size of 512, and fine-tune for 10 epochs. The learning rate is set to 3×10^{-5} for all parameters and follows a cosine-annealing schedule [27] with 500 warm-up steps. No data augmentation is applied apart from the necessary resizing and cropping, following the training of CLIP. We split out a validation set of 10240 samples from the ImageNet training set to perform early stopping and model selection based on the validation accuracy. β in Eq. (2) is fixed to be 30 for all experiments. Additional details are in the appendix.

4.2. Masking and Refilling Strategies

In this section, we validate the effectiveness of using masked images for improving OOD robustness, and investigate the different masking and refilling strategies discussed in Sec. 3.2. Concretely, we consider three masking strategies: (1) *random-mask*: randomly select a fixed proportion of patches; (2) *context-mask*: masking the patches with CAM score lower than the threshold; (3) *object-mask*: masking the patches with CAM score higher than the threshold. The masking rate for random-mask is selected from $\{0.25, 0.5, 0.75\}$, and the CAM score threshold for the latter two strategies is selected from $\{0.3, 0.4, 0.5, 0.6\}$. The refilling strategies include: (1) *no-fill*: baseline strategy that drops the masked patches without refilling; (2) *single-fill*: refill with patches from one other image; (3) *multi-fill*: refill with patches from multiple other images. We test all the combinations of masking and refilling strategies, and compare them with the baseline where no masking is applied (*i.e.*, $x_{cf} = x$ in Eq. (2)).

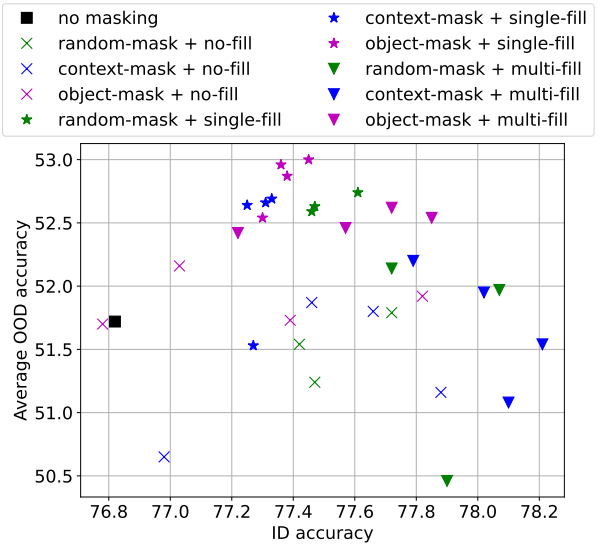

Figure 6. ID and average OOD accuracy of our approach using different masking and refilling strategies. For each combination of masking and refilling strategies, we plot multiple results based on different settings of hyper-parameters (*i.e.*, masking rate or CAM score threshold). Best viewed in color.

In Fig. 6, we plot the average OOD accuracy against ID accuracy for each model trained with the above strategies. Besides, for each combination of masking and refilling strategies, we select the model with highest validation accuracy, and report their results on each OOD dataset in Tab. 1. Based on these results, we have the following conclusions. (1) Most of the combinations of masking and refilling strategies achieve better ID-OOD trade-off than the no-masking baseline, which suggests the effectiveness of image masking in our proposed approach. (2) Refilling masked images with patches from other images (*i.e.*, single-fill or multi-fill) is better than solely dropping the masked patches. (3) Comparing the two proposed refilling strategies, single-fill generally results in better OOD accuracy, while multi-fill may

Method	Masking	Refilling	IN	IN-V2	IN-R	IN-Sketch	ObjectNet	IN-A	OOD avg.
Zero-shot [34]	/	/	63.4	55.9	69.3	42.3	44.5	31.4	48.7
Vanilla fine-tuning	/	/	75.9	64.7	57.0	39.8	39.5	20.0	44.2
Ours	no masking	/	76.9	66.5	69.2	45.6	45.3	29.8	51.3
Ours	random-mask	no-fill	77.5	66.9	66.4	45.7	46.5	30.8	51.2
	context-mask		77.8	67.4	66.7	45.6	45.9	30.0	51.1
	object-mask		77.7	67.1	67.6	46.2	46.8	31.5	51.9
Ours	random-mask	single-fill	77.6	67.1	69.0	46.4	47.8	33.4	52.7
	context-mask		77.2	66.9	68.8	46.5	47.8	32.9	52.6
	object-mask		77.5	67.1	**69.7**	**46.9**	**48.0**	**33.8**	**53.1**
Ours	random-mask	multi-fill	78.0	67.4	67.4	46.1	46.7	31.9	51.9
	context-mask		**78.2**	67.4	66.5	45.5	45.9	30.0	51.1
	object-mask		77.9	**67.7**	68.1	46.6	47.5	33.0	52.6

Table 1. Comparison of different masking and refilling strategies. Accuracy on ImageNet (IN) and the five OOD datasets are reported. *OOD avg.* is the average OOD accuracy on the five OOD datasets. The best accuracy is **bold-faced**, and the second best accuracy is underlined. Results are averaged over three runs with different seeds.

Threshold	0.7	0.6	0.5	0.4	0.3	0.2
Image MR	0.06	0.10	0.15	0.25	0.39	0.55
Object MR	0.17	0.23	0.33	0.48	0.65	0.80
IoU	0.15	0.21	0.28	0.38	0.45	0.48

Table 2. Validation of CAM-based object masking with different thresholds. Image masking rate (MR): masking rate of the whole image. Object masking rate (MR): masking rate concerning the object area. IoU: Intersection over Union regrading the object. Averaged over a subset of masked samples constructed in training.

yield better ID accuracy. (4) Comparing the masking strategies, object-mask is generally better than random-mask and context-mask in terms of OOD accuracy.

The superiority of object-mask can be explained following our analysis in Sec. 3.2.2. Concretely, while both masking strategies are theoretically valid, context-mask is more dependent on refilling, since it supposes that the refilled context can effectively lead to contradictions. This may not be satisfied by the proposed refilling strategies, as they are not aware of the content of the patches taken from other images. Hence, context-mask can be less effective in practice.

In addition, to verify that our CAM-based object masking results in superior OOD accuracy due to the effective masking of objects, we calculate the average masking rates of the objects during training. Specifically, we experiment on the object-mask and single-fill strategy with different CAM score thresholds. To obtain the masking rate of the object for an ImageNet image, we take the pixel-level mask produced by a pre-trained segmentation model [11] as the approximation of ground truth. In addition, we compute the masking rate of the whole image and the Intersection over Union (IoU) regrading the object. Details are provided in the appendix. As shown in Tab. 2, object masking rate is significantly higher than the masking rate of the whole image, which suggests that our CAM-based masking strategy mainly masks on the objects. Besides, it is shown that as the threshold for masking decreases, both the object masking rate and IoU increase. Hence, the CAM-based object masking is empirically sound.

4.3. Comparison with Existing Approaches

We compare our method with three existing approaches, namely LP-FT [23], WiSE-FT [46] and Model Soup [45], which also aim to improve both ID and OOD accuracy of fine-tuning. For WiSE-FT, we report the results with the default mixing coefficient, *i.e.*, $\alpha = 0.5$. For Model Soup, we take the *uniform soup* as the representative method, which achieves the best average OOD accuracy on CLIP fine-tuning as reported in [45]. For our approach, we consider both refilling strategies, and adopt the object masking with the CAM score threshold that yields highest validation accuracy. The results are shown in Tab. 3. **First**, our approach surpasses the previous approaches in terms of average OOD accuracy. **Second**, our approach achieves better ID accuracy than other approaches except for Model Soup. Note that Model Soup takes the ensemble of many fine-tuned models with different hyper-parameters, including those trained with strong data augmentation. Conversely, we do not dive into hyper-parameters tuning and do not use data augmentation in our experiments. **Third**, compared with WiSE-FT, our approach achieves superior performance on ObjectNet and ImageNet-A, but is inferior on ImageNet-R and ImageNet-Sketch, which suggests that the two approaches may improve different perspectives of robustness. This motivates us to consider the integration of WiSE-FT and our approach.

4.4. Integrating WiSE-FT

WiSE-FT ensemble the weights of the zero-shot model and the vanilla fine-tuned model, and the ensemble model may achieve better accuracy on both ID and OOD data. Concretely, the weight ensemble is computed by $\theta_e = (1-\alpha)\cdot\theta_0 + \alpha\theta_1$, where θ_0 is the weights of the zero-

Model	Method	IN	IN-V2	IN-R	IN-Sketch	ObjectNet	IN-A	OOD avg.
CLIP ViT-B/32	Zero-shot [34]	63.4	55.9	69.3	42.3	44.5	31.4	48.7
	Vanilla fine-tuning	75.9	64.7	57.0	39.8	39.5	20.0	44.2
	WiSE-FT† [46]	76.6	66.6	**70.2**	47.1	46.3	31.9	52.4
	Uniform soup‡ [45]	**80.0**	**68.6**	66.6	**47.7**	46.1	29.2	51.6
	Ours (multi-fill)	77.9	67.7	68.1	46.6	47.5	33.0	52.6
	Ours (single-fill)	77.5	67.1	69.7	46.9	**48.0**	**33.8**	**53.1**
CLIP ViT-B/16	Zero-shot [34]	68.3	61.9	77.6	48.3	54.0	50.1	58.4
	Vanilla fine-tuning	80.7	70.4	64.0	45.1	49.1	35.2	52.8
	LP-FT [23]	81.7	71.6	72.9	48.4	/	49.1	/
	WiSE-FT [46]	81.7	72.8	**78.7**	**53.9**	57.3	52.2	63.0
	Ours (multi-fill)	**82.5**	73.4	76.4	52.7	56.8	52.0	62.3
	Ours (single-fill)	82.4	**73.4**	78.1	53.4	**57.9**	**53.5**	**63.3**

Table 3. Accuracy of different methods for fine-tuning CLIP models on ImageNet (IN). *OOD avg.* is the average OOD accuracy on the five OOD datasets. The best accuracy is **bold-faced**, and the second best accuracy is underlined. Our results are averaged over three runs with different seeds. (†: our implementation. ‡: our evaluation on the official model.)

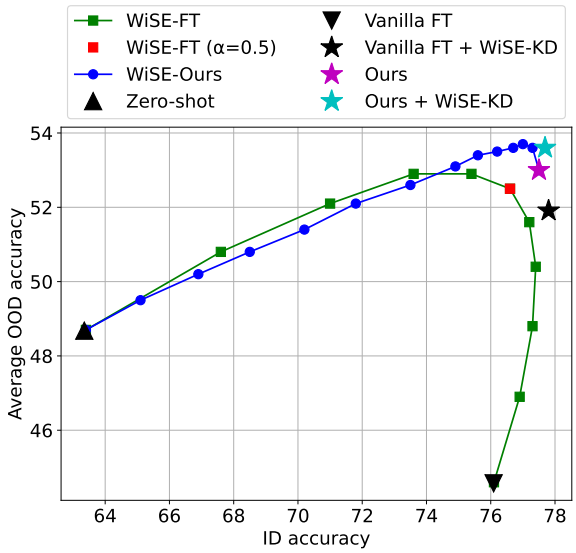

Figure 7. Integration of WiSE-FT and our approach. WiSE-Ours: weight-space ensembles of the zero-shot model and our fine-tuned model. WiSE-KD: knowledge distillation from the WiSE-FT model ($\alpha = 0.5$). Vanilla FT: vanilla fine-tuning. For WiSE-FT and WiSE-Ours, each point corresponds to a value of $\alpha \in [0, 1]$.

shot model, and θ_1 is the weights of the fine-tuned model. This weight-space ensemble method is agnostic of the fine-tuning method in form. Therefore, we may expect that the ensemble of the zero-shot model and the model fine-tuned with our method can achieve even better results. However, as shown in Fig. 7, such direct integration with WiSE-FT may not improve ID and OOD accuracy simultaneously. Specifically, only for α close to 1, the resulting model may achieve better average OOD accuracy, but still at the cost of lower ID accuracy. This is different from the observations of WiSE-FT that using a medium value of α (*e.g.*, 0.5) usually yields near-optimal results and surpass the fine-tuned model on both ID and OOD accuracy [46]. It is suggested that our proposed method may produce a significantly different model as compared with vanilla fine-tuning. In other words, our fine-tuned weights may leave the basin of the loss landscape where the weights of most vanilla fine-tuned models and the pre-trained model lie in [31, 45, 46].

Another way to integrate WiSE-FT into our method is to utilize the ensemble model produced by WiSE-FT to guide the fine-tuning model via knowledge distillation (KD) [17]. For brevity, we name this method as "WiSE-KD". Specifically, we consider using the WiSE-FT model as another teacher model, and add the vanilla knowledge distillation loss [17] to our training objective Eq. (2). Details are presented in appendix. As presented in Fig. 7, applying WiSE-KD to our approach slightly improves both ID and OOD accuracy. As a comparison, while applying WiSE-KD to vanilla fine-tuning results in significantly better performance, it is still inferior to our approach, especially for OOD accuracy. This again validates the effectiveness of our approach in improving the OOD robustness in fine-tuning.

5. Conclusion and Limitation

In this paper, we analyze the issue of robustness degradation in fine-tuning from a causal perspective, and find that masked samples can be effective counterfactual samples for improving the OOD robustness of fine-tuning. Experiments suggest that our approach surpass previous methods on the OOD performance with competitive ID performance.

As discussed in Sec. 4.2, the proposed refilling strategies are unaware of the content of the patches to fill, which limit the effectiveness of the resulting counterfactual samples. Besides, the feature-based distillation may not regularize the learning of the classification head of the fine-tuning model. Future works can devote to better refilling and distillation methods conforming to the causal modeling.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020. 2

[2] Martin Arjovsky. *Out of distribution generalization in machine learning*. PhD thesis, New York University, 2020. 2

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 2

[5] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019. 1, 5, 11

[6] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 4

[7] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*, 2019. 2

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1, 2

[9] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 2, 5, 6

[10] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020. 2

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 7, 11, 12

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 5, 6

[13] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5078–5088, 2022. 2

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 5

[15] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1, 5, 11

[16] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 1, 5, 11

[17] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 8, 11

[18] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *International Conference on Machine Learning*, pages 4555–4562. PMLR, 2021. 2, 3

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2, 5

[20] Nan Rosemary Ke, Aniket Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Christopher Pal. Systematic evaluation of causal discovery in visual model based reinforcement learning. *arXiv preprint arXiv:2107.00848*, 2021. 2, 3

[21] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020. 2

[22] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 2

[23] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2021. 1, 2, 3, 7, 8

[24] Ananya Kumar, Aditi Raghunathan, Tengyu Ma, and Percy Liang. Calibrated ensembles: A simple way to mitigate id-ood accuracy tradeoffs. In *NeurIPS 2021 Workshop on*

Distribution Shifts: Connecting Methods and Applications, 2021. 2

[25] 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. 11

[26] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34:6155–6170, 2021. 2

[27] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. 2016. 6, 11

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6, 11

[29] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021. 2, 3

[30] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021. 1

[31] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020. 8

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 11

[33] Judea Pearl. *Causality*. Cambridge university press, 2009. 2, 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5, 6, 7, 8, 11

[35] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1, 5, 11

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5, 11

[37] Nino Scherrer, Anirudh Goyal, Stefan Bauer, Yoshua Bengio, and Nan Rosemary Ke. On the generalization and adaption performance of causal models. *arXiv preprint arXiv:2206.04620*, 2022. 2, 3

[38] Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109:612–634, 2021. 2, 3

[39] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019. 3

[40] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020. 1

[41] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2018. 11

[42] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 5, 11

[43] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 2

[44] Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality? *arXiv preprint arXiv:2206.10591*, 2022. 1

[45] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998. PMLR, 2022. 1, 2, 7, 8, 11

[46] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 1, 2, 5, 6, 7, 8, 11

[47] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2

[48] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3

[49] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 2921–2929, 2016. 2, 5

A. Experiment Details

A.1. Training Routines

For fine-tuning on ImageNet (including vanilla fine-tuning and our approach), we use the AdamW optimizer [28] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 0.1 and gradient clipping at ℓ_2-norm 1. We use a batch size of 512, and fine-tune for 10 epochs. The learning rate is set to 3×10^{-5} for all parameters and follows a cosine-annealing schedule [27] with 500 warm-up steps. For both training and testing, we resize and center-crop the images to the size of 224×224, and no data augmentation is applied. Besides, different from WiSE-FT [46], we do not use label smoothing.

A.2. Validation of CAM-based Object Masking

In Sec. 4.2, to verify that our CAM-based object masking can effectively mask the patches that cover the main object, we report the average object masking rate and IoU during training with different CAM score thresholds. Since we do not have the ground truth of the masks of main objects for ImageNet, we approximate it by the prediction of Mask2Former [11], a segmentation model pre-trained on COCO [25] (the specific model is reported in Appendix B). We select three super-classes defined in Restricted ImageNet [41] that can be recognized by the segmentation model, *i.e.*, Dog, Cat and Bird, which cover 144 ImageNet classes in total. For each training image of these classes, we obtain the pixel-level segmentation mask M_{seg} corresponding to the super-class, and compare it with our patch-level CAM-based mask, which is translated to a pixel-level mask M_{CAM} according to the correspondence between patches and pixels.

The metrics in Tab. 2 in the main text are defined as follows. Formally, a mask M of an image I is defined as a subset of the pixels. Let $n(\cdot)$ denote the number of pixels in a mask or an image. Then, the metrics are defined as:

- Image masking rate: $\dfrac{n(M_{\text{CAM}})}{n(I)}$;
- Object masking rate: $\dfrac{n(M_{\text{CAM}} \cap M_{seg})}{n(M_{seg})}$;
- IoU: $\dfrac{n(M_{\text{CAM}} \cap M_{seg})}{n(M_{\text{CAM}} \cup M_{seg})}$.

A.3. WiSE-KD

In Sec. 4.4, we consider using the WiSE-FT [46] model as a teacher model, and add the vanilla knowledge distillation loss [17] to our training objective, , *i.e.*,

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{CE}}(g(f(x))) + \gamma \mathcal{L}_{\text{KL}}(g(f(x)), g_e(f_e(x))) \\ + \beta \mathcal{L}_{\text{MSE}}(\hat{f}(x_{cf}), f(x_{cf})),\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{KL}}$ is the Kullback-Leibler divergence loss, and f_e and g_e are the encoder and classification head of the ensemble model produced by WiSE-FT, correspondingly. We set $\gamma = 1$, and use the WiSE-FT model with $\alpha = 0.5$. The temperature of the vanilla distillation is 10.

B. Use of Existing Assets

Datasets. In this paper, we utilize the following existing benchmark datasets without modification or repackaging:

- ImageNet [36] (https://www.image-net.org/)
- ImageNet-V2 [35] (https://github.com/modestyachts/ImageNetV2)
- ImageNet-R [15] (https://github.com/hendrycks/imagenet-r)
- ImageNet-Sketch [42] (https://github.com/HaohanWang/ImageNet-Sketch)
- ObjectNet [5] (https://objectnet.dev/)
- ImageNet-A [16] (https://github.com/hendrycks/natural-adv-examples)

In our experiments, we select the hyper-parameters based on validation accuracy on ImageNet, and use the other datasets solely for robustness evaluation. For ObjectNet, we follow the official guidance to remove the red borders of the images before other preprocessing steps in evaluation.

Code and pre-trained model weights. The experiments in this paper are based on the code and pre-trained model weights provided by the following packages or GitHub repositories:

- PyTorch [32] (https://github.com/pytorch/pytorch)
- CLIP [34] (https://github.com/openai/CLIP)
- WiSE-FT [46] (https://github.com/mlfoundations/wise-ft)
- Model Soup [45] (https://github.com/mlfoundations/model-soups/issues/1): we use the pre-trained weights of uniform soup provided by the authors in an issue.

- Mask2Former [11] (https://github.com/facebookresearch/Mask2Former/blob/main/MODEL_ZOO.md): we use the pre-trained model with ID 48558700_7.