Masked Images Are Counterfactual Samples for Robust Fine-tuning (Appendix)

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

A. Experiment Details

A.1. Training Routines

For fine-tuning on ImageNet via vanilla fine-tuning or our approach, we use the AdamW optimizer [8] 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 [7] 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 [16], 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 [2], a segmentation model pre-trained on COCO [6] (the specific model is reported in Appendix B). We select three super-classes defined in Restricted ImageNet [13] 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 [16] model as a teacher model, and add the vanilla knowledge distillation loss [5] to our training objective, *i.e.*,

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{CE}}(g(f(x)), y) + \gamma\mathcal{L}_{\text{KL}}(g(f(x)), g_e(f_e(x))) \\ +\beta\mathcal{L}_{\text{MSE}}(\hat{f}(x_{cf}), f(x_{cf})),\end{aligned} \tag{1}$$

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 knowledge distillation is 10.

B. Use of Existing Assets

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

- ImageNet [12] (https://www.image-net.org/)
- ImageNet-V2 [11] (https://github.com/modestyachts/ImageNetV2)
- ImageNet-R [3] (https://github.com/hendrycks/imagenet-r)
- ImageNet-Sketch [14] (https://github.com/HaohanWang/ImageNet-Sketch)
- ObjectNet [1] (https://objectnet.dev/)

*Corresponding author.

- ImageNet-A [4] (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 [9] (https://github.com/pytorch/pytorch)
- CLIP [10] (https://github.com/openai/CLIP)
- WiSE-FT [16] (https://github.com/mlfoundations/wise-ft)
- Model Soup [15] (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 [2] (https://github.com/facebookresearch/Mask2Former/blob/main/MODEL_ZOO.md): we use the pre-trained model with ID 48558700_7.

References

[1] 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

[2] 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. 1, 2

[3] 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

[4] 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. 2

[5] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1

[6] 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. 1

[7] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. 2016. 1

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 1

[9] 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. 2

[10] 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. 2

[11] 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

[12] 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. 1

[13] 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. 1

[14] 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. 1

[15] 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. 2

[16] 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