Masked Images Are Counterfactual Samples for Robust Fine-tuning

Yao Xiao Ziyi Tang Pengxu Wei* Cong Liu Liang Lin

*Corresponding author

Sun Yat-sen University

Poster: THU-AM-363

Project/Code: https://github.com/Coxy7/robust-finetuning/

Robustness to distribution shift

In-distribution (ID)

Out-of-distribution (OOD)

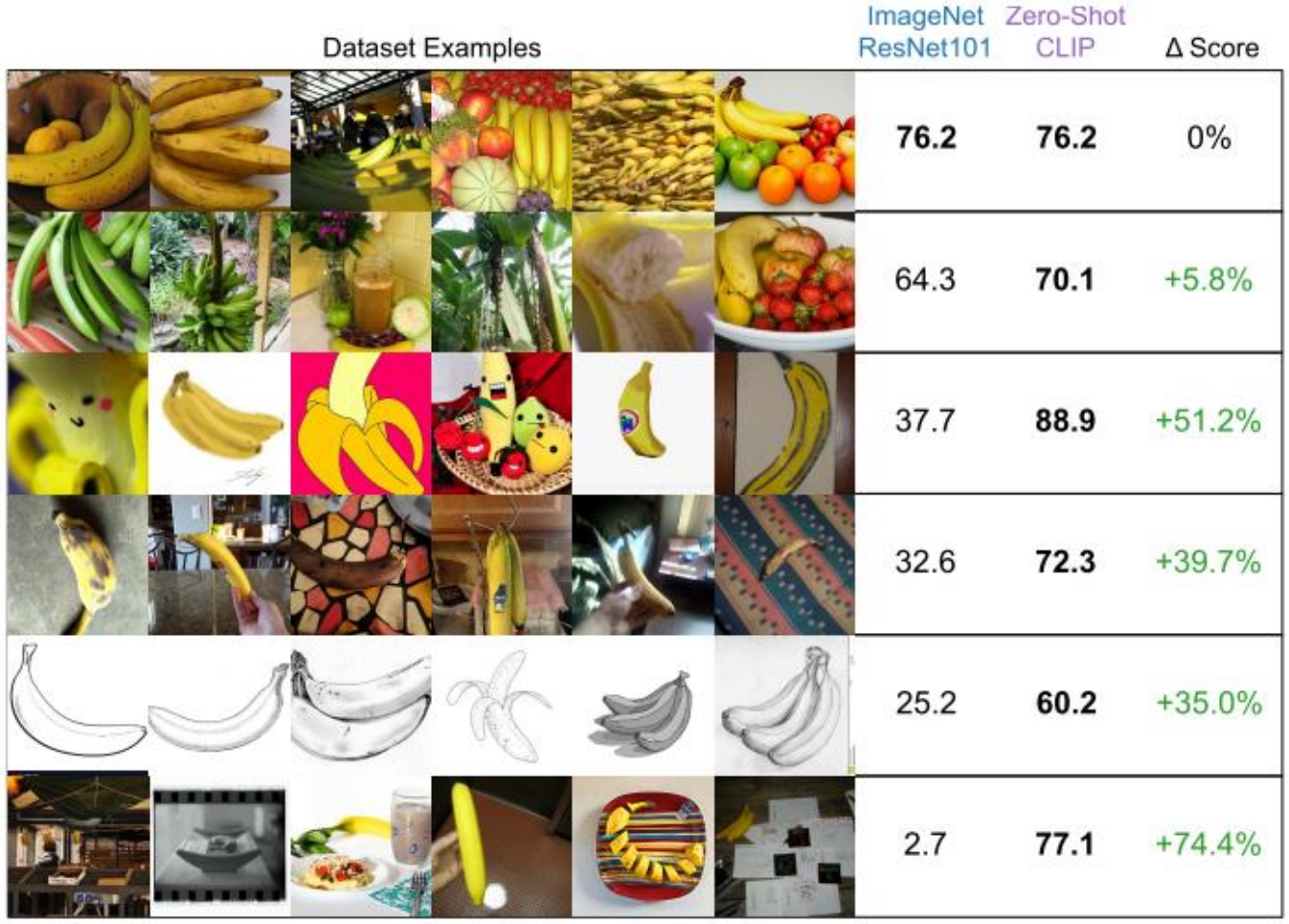


Figure source: Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML, 2021.

Fine-tuning can reduce the robustness

- Trade-off between ID and OOD performance

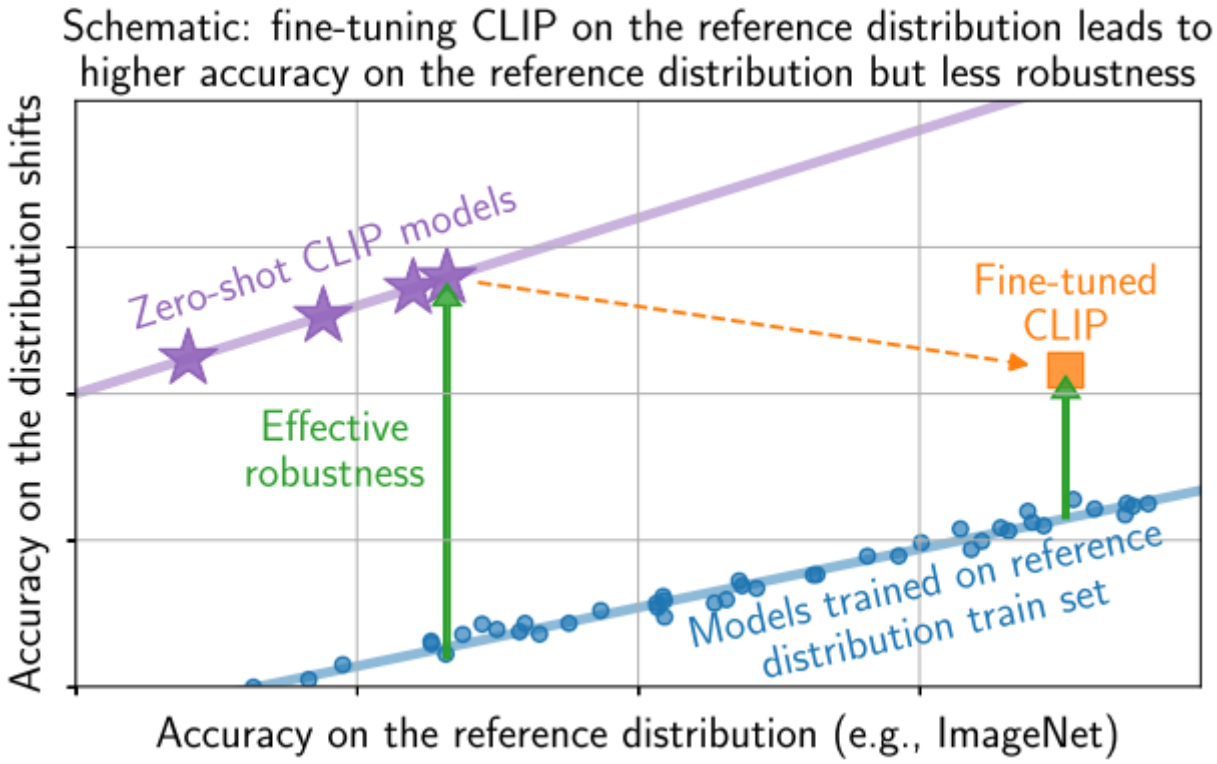
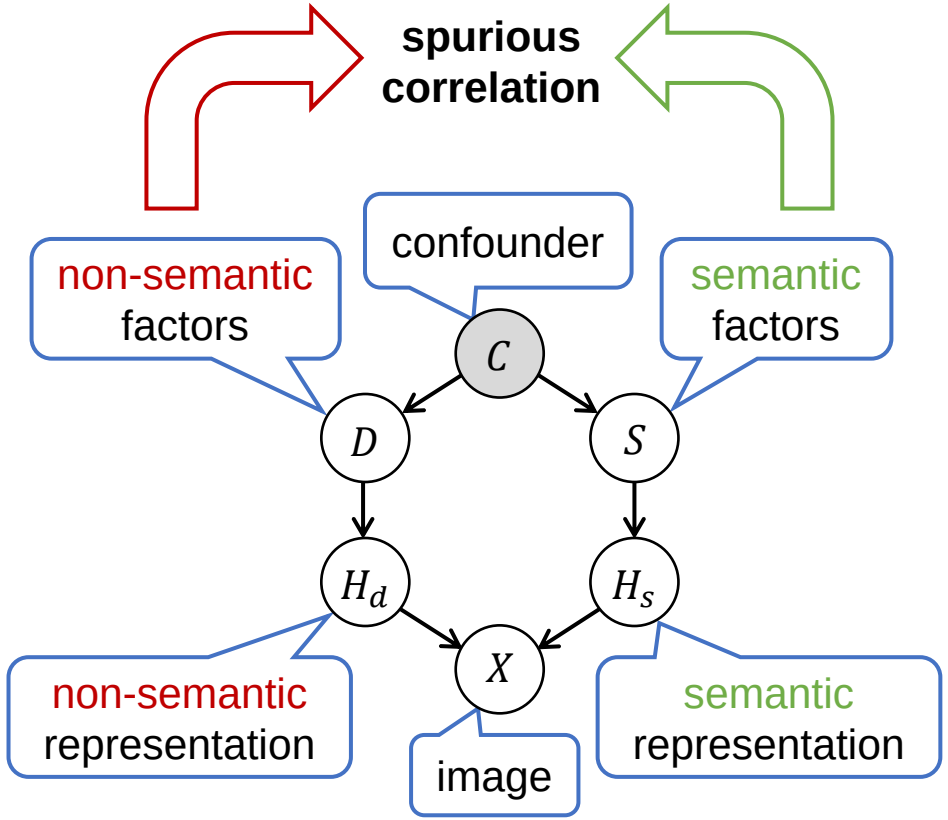
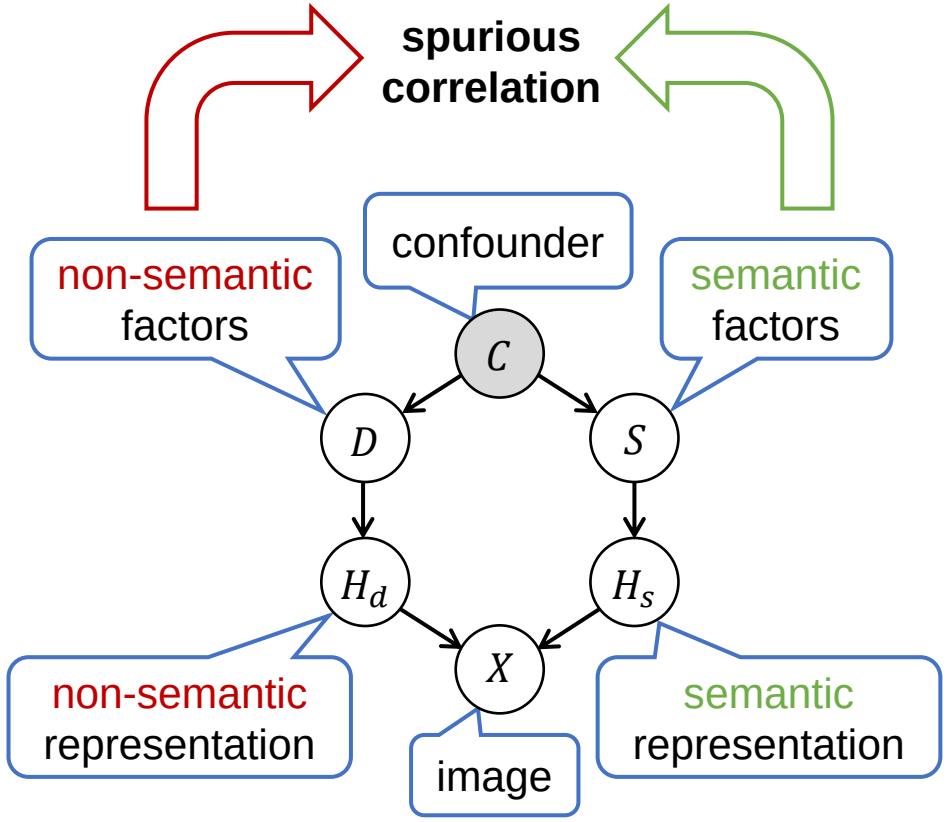


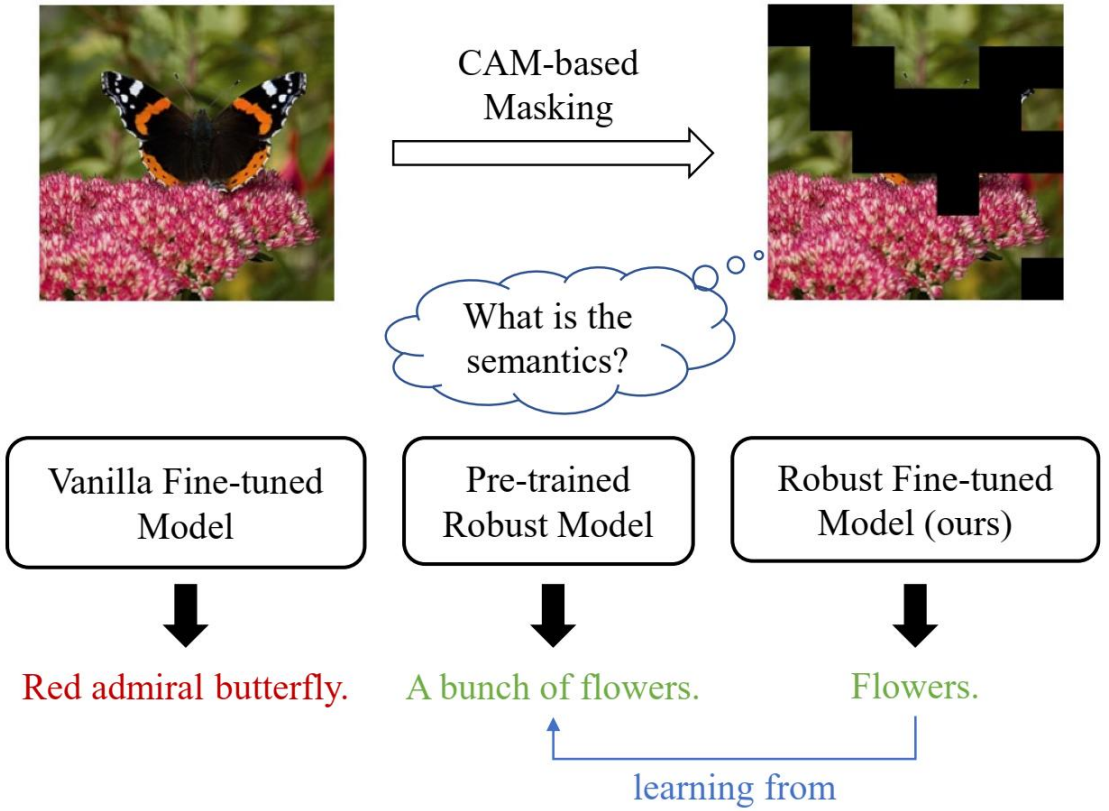
Figure source: Wortsman, Mitchell, et al. "Robust fine-tuning of zero-shot models." CVPR, 2022.

Overview


spurious correlation
confounder
non-semantic factors
semantic factors
C
D
S
H_d
H_s
X
non-semantic representation
semantic representation
image

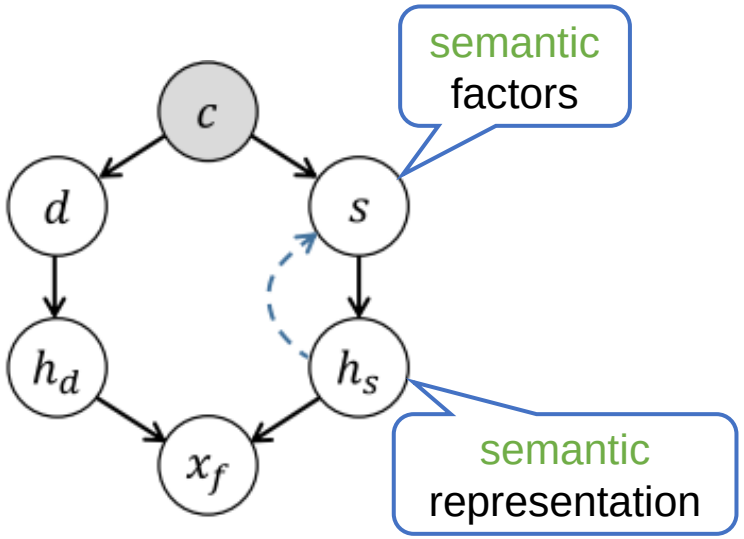
Overview


spurious correlation
confounder
non-semantic factors
semantic factors
C
D
S
H_d
H_s
X
non-semantic representation
semantic representation
image

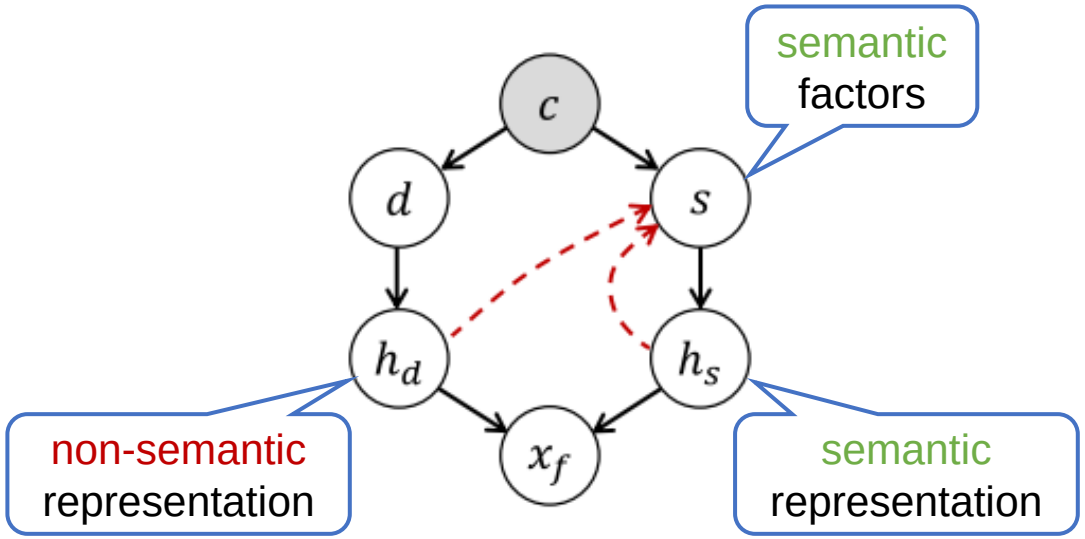

CAM-based Masking
What is the semantics?
Vanilla Fine-tuned Model
Pre-trained Robust Model
Robust Fine-tuned Model (ours)
Red admiral butterfly.
A bunch of flowers.
Flowers.
learning from

Fine-tuning can reduce the robustness: a causal perspective

- Fine-tuned models tend to rely on both semantic & non-semantic representations (h_s, h_d) for the prediction of image semantics s



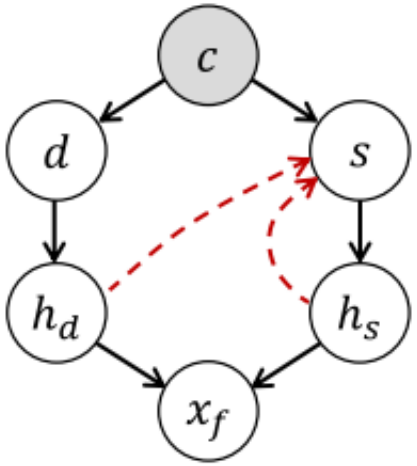
Pre-trained models:
Predict s from h_s



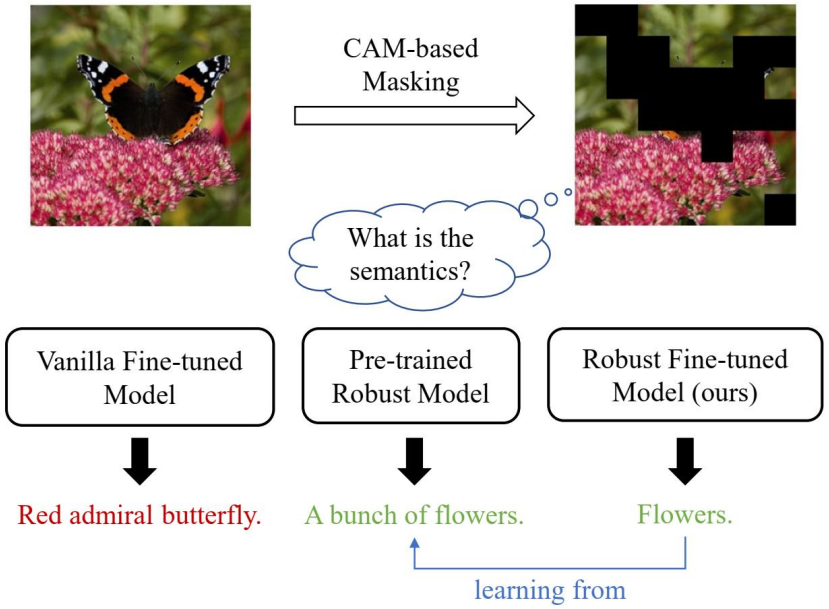
Fine-tuned models:
Predict s from h_s & h_d due to the *spurious correlation*

Fine-tuning can reduce the robustness: a causal perspective

- Fine-tuned models tend to rely on both semantic & non-semantic representations (h_s, h_d) for the prediction of image semantics s
- The correlation between h_d and s is unstable under distribution shift

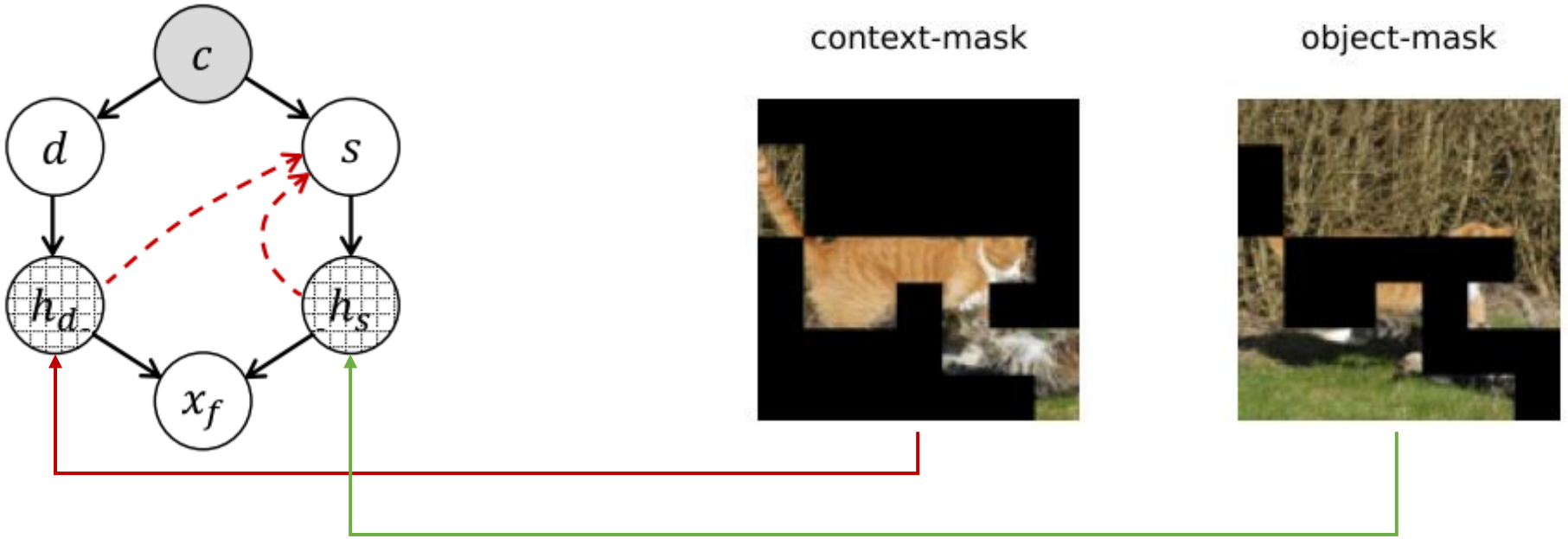


Fine-tuned models:
Predict s from h_s & h_d due to the spurious correlation



Masked images as counterfactual samples

- Masking based on Class Activation Map (CAM):
  - Context-mask: h_d
  - Object-mask: h_s



Robust fine-tuning with counterfactual samples

- Distillation with the pre-trained model based on masked images x_{cf}

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}\big(g\big(f(x)\big), y\big) + \beta \mathcal{L}_{\mathrm{MSE}}\left(\hat{f}\big(x_{cf}\big), f\big(x_{cf}\big)\right)$$

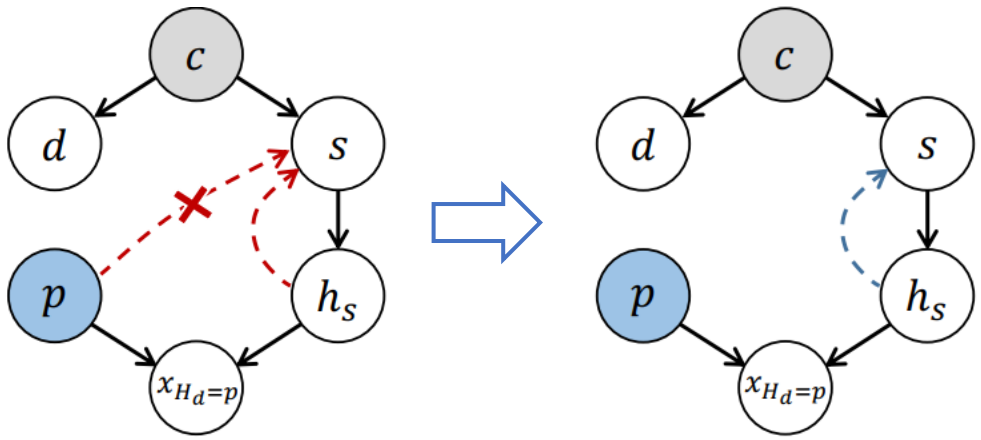
x: factual sample; x_{cf}: counterfactual sample; y: label
f, $\hat{f}$: backbone of fine-tuning/pre-trained model; g: classifier

Robust fine-tuning with counterfactual samples

- Distillation with the pre-trained model based on masked images x_{cf}

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

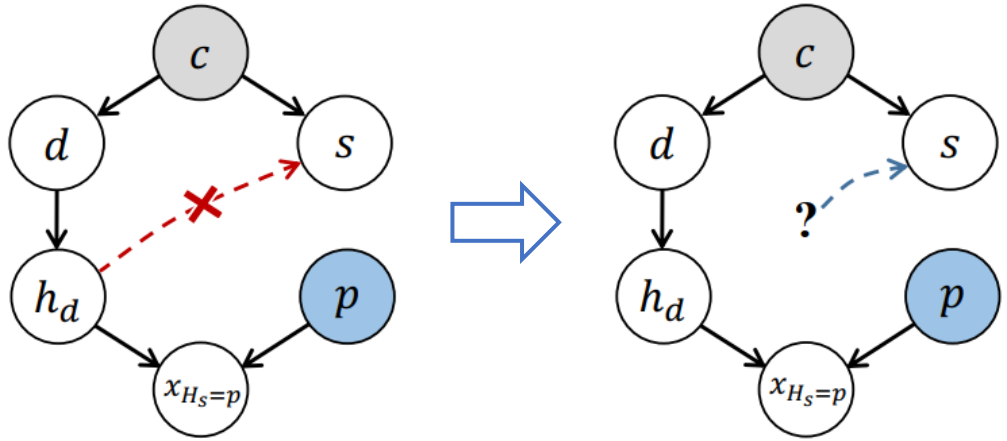
Context-mask



Fine-tuned model

Pre-trained model

Object-mask

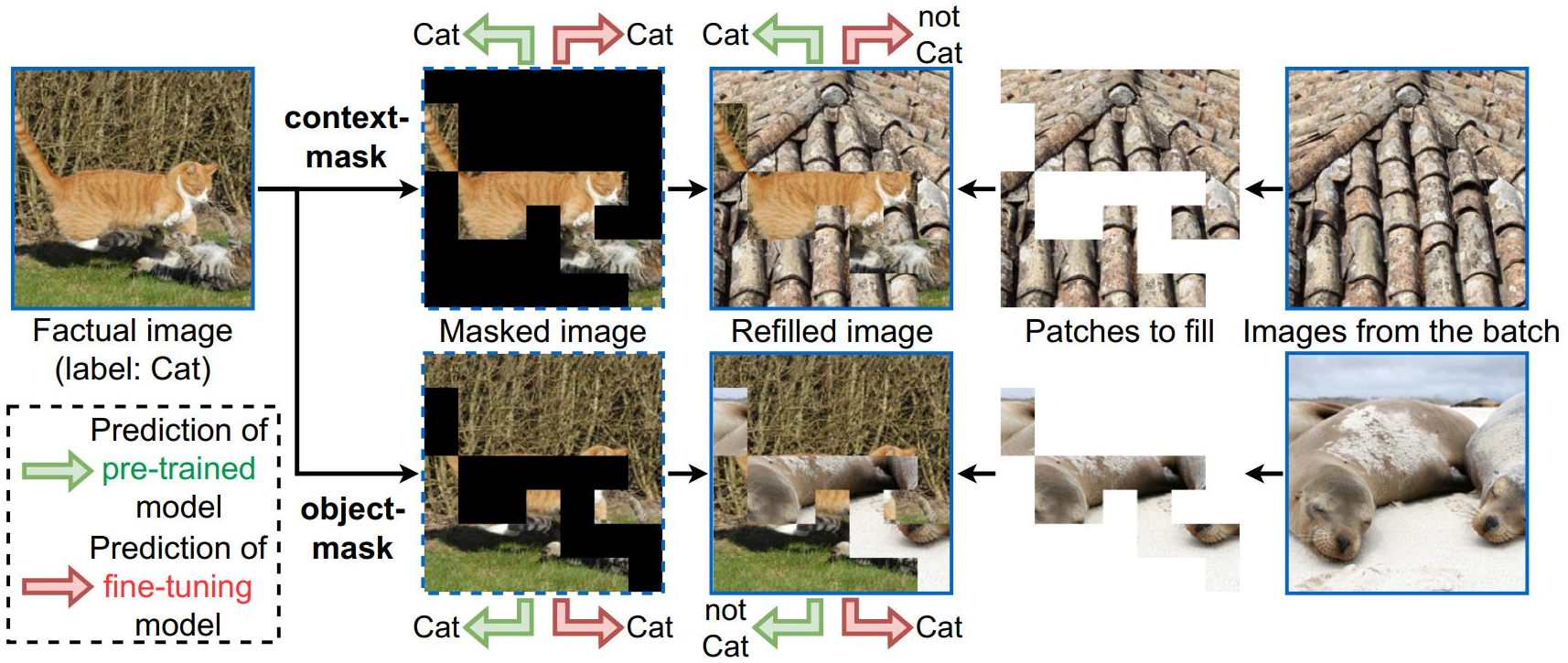


Fine-tuned model

Pre-trained model

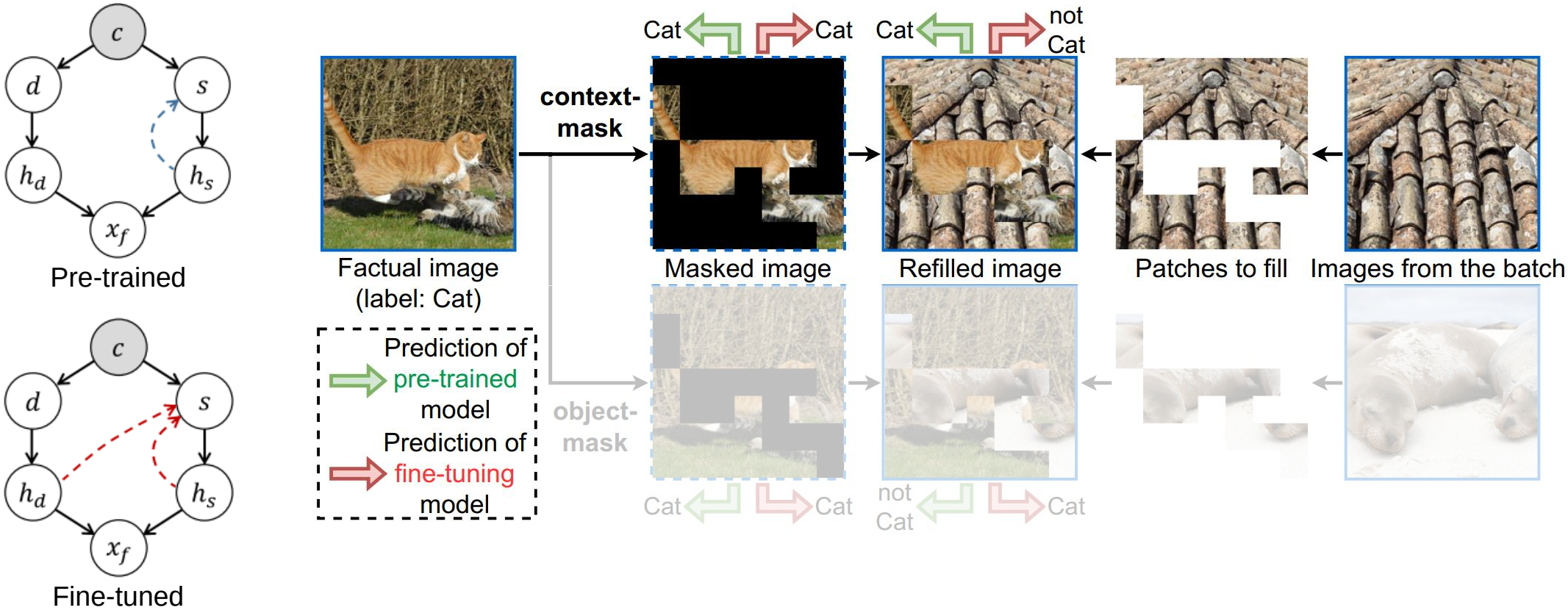
Robust fine-tuning with counterfactual samples

- Refilling
  - To enlarge the **disagreement** between fine-tuning and pre-trained model on x_{cf}



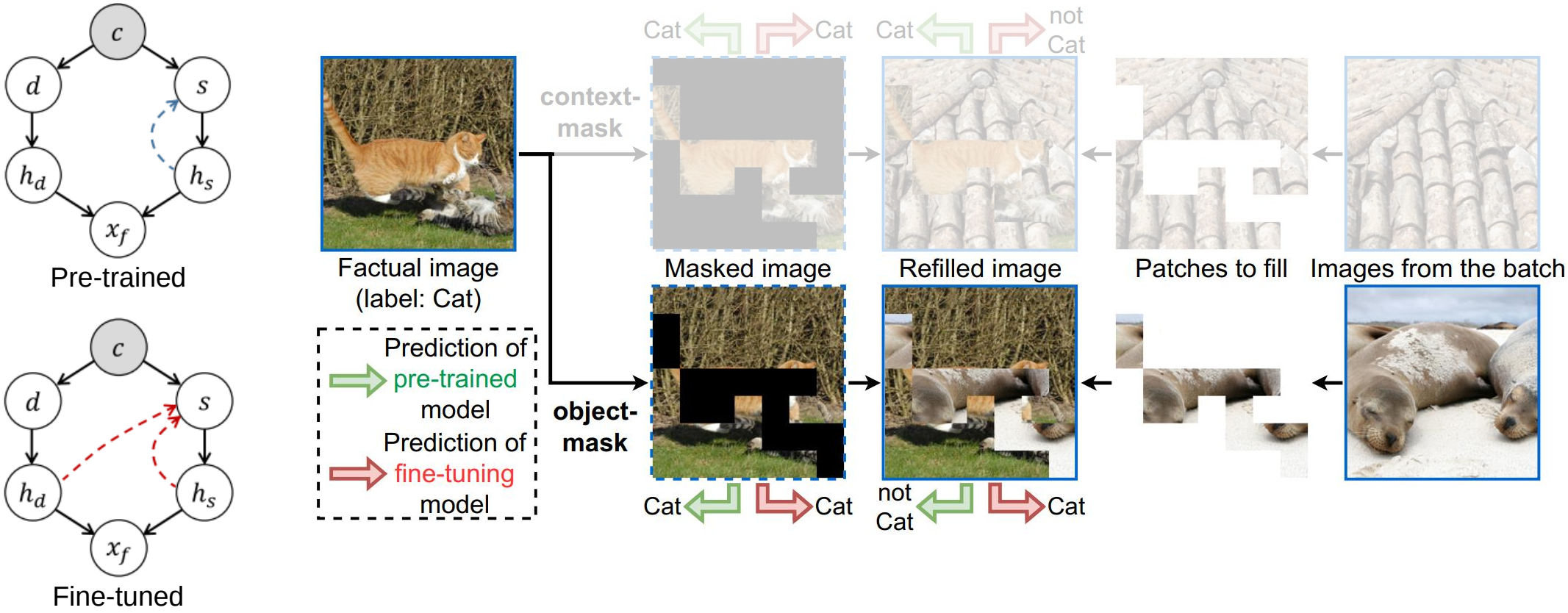
Robust fine-tuning with counterfactual samples

- Refilling
  - To enlarge the **disagreement** between fine-tuning and pre-trained model on x_{cf}

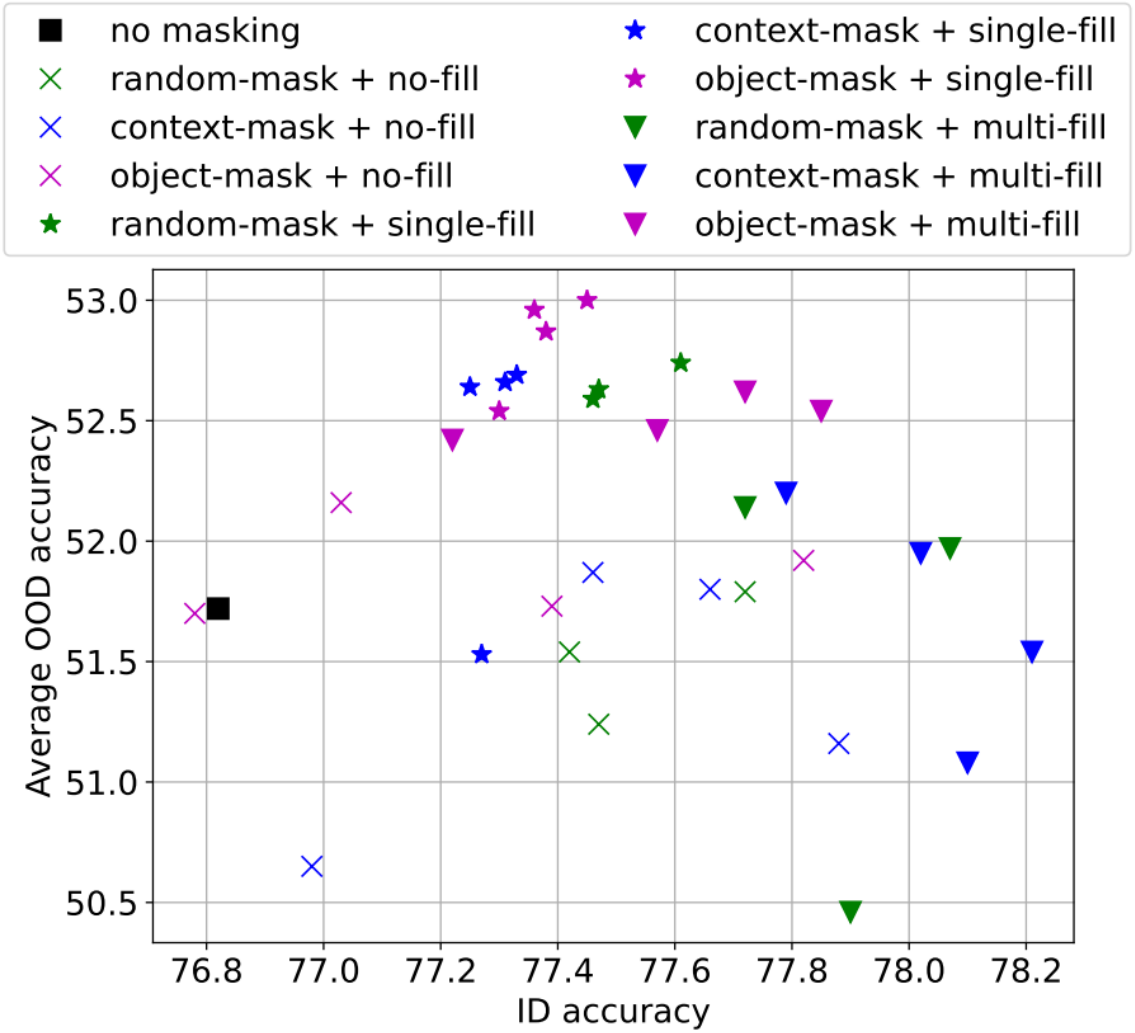


Robust fine-tuning with counterfactual samples

- Refilling
  - To enlarge the **disagreement** between fine-tuning and pre-trained model on x_{cf}



Validation of masking and refilling strategies



Hyper-parameters:

Random-mask:

- Masking rate in {0.25, 0.5, 0.75}

Context/Object-mask:

- CAM score threshold in {0.3, 0.4, 0.5, 0.6}

Conclusions:

1. Masking > no masking
2. Refilling > no refilling
3. Object-mask > random/context mask

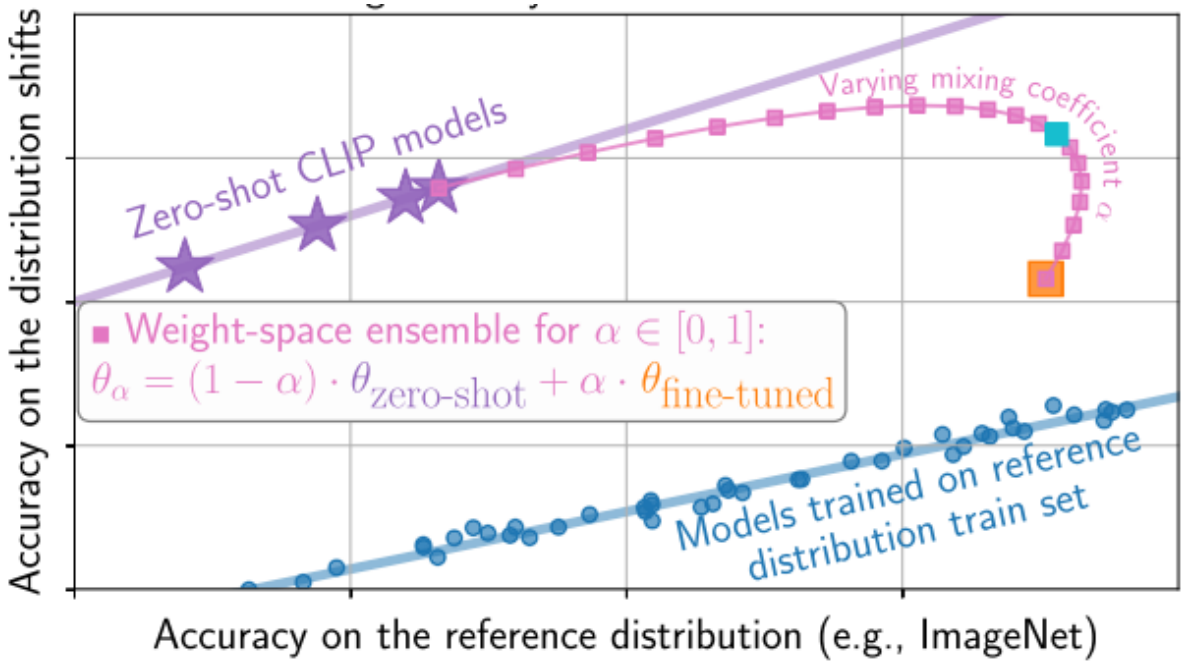
Comparison with existing methods

In-distribution (ID) | Out-of-distribution (OOD)

Model	Method	IN	IN-V2	IN-R	IN-Sketch	ObjectNet	IN-A	OOD avg.
CLIP ViT-B/32	Zero-shot [32]	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† [43]	76.6	66.6	**70.2**	47.1	46.3	31.9	52.4
	Uniform soup‡ [42]	**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 [32]	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 [43]	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**

- The proposed method is particularly effective when objects are shown in **unusual contexts** (ObjectNet, ImageNet-A)

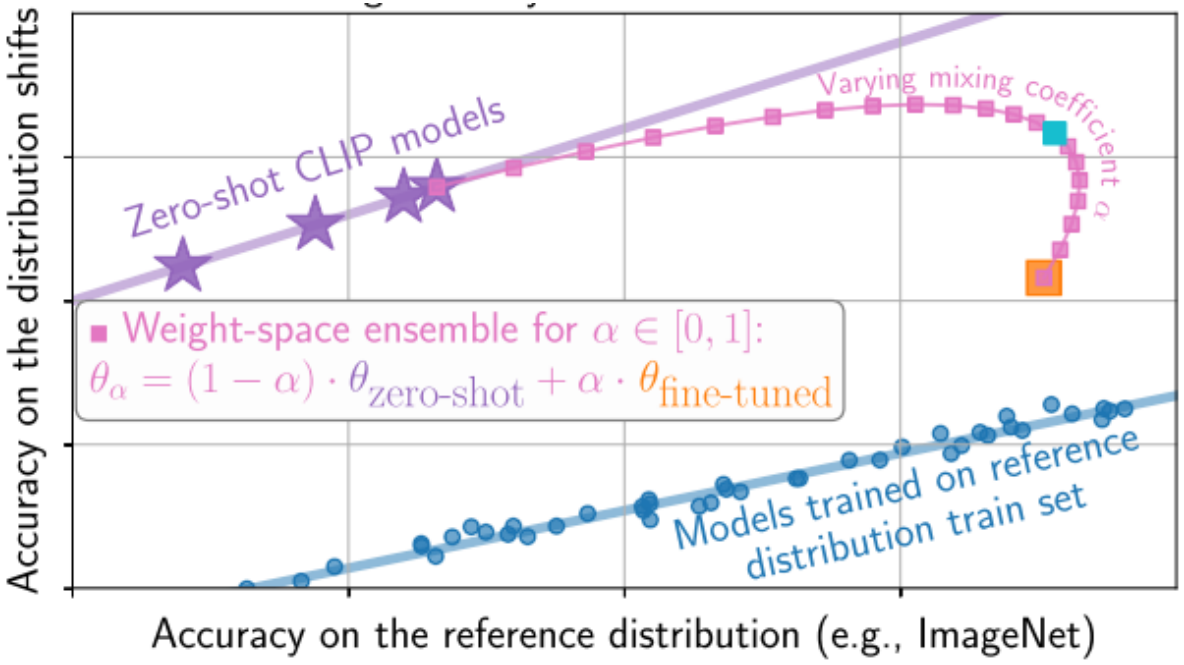
Discussion: WiSE-FT [1]



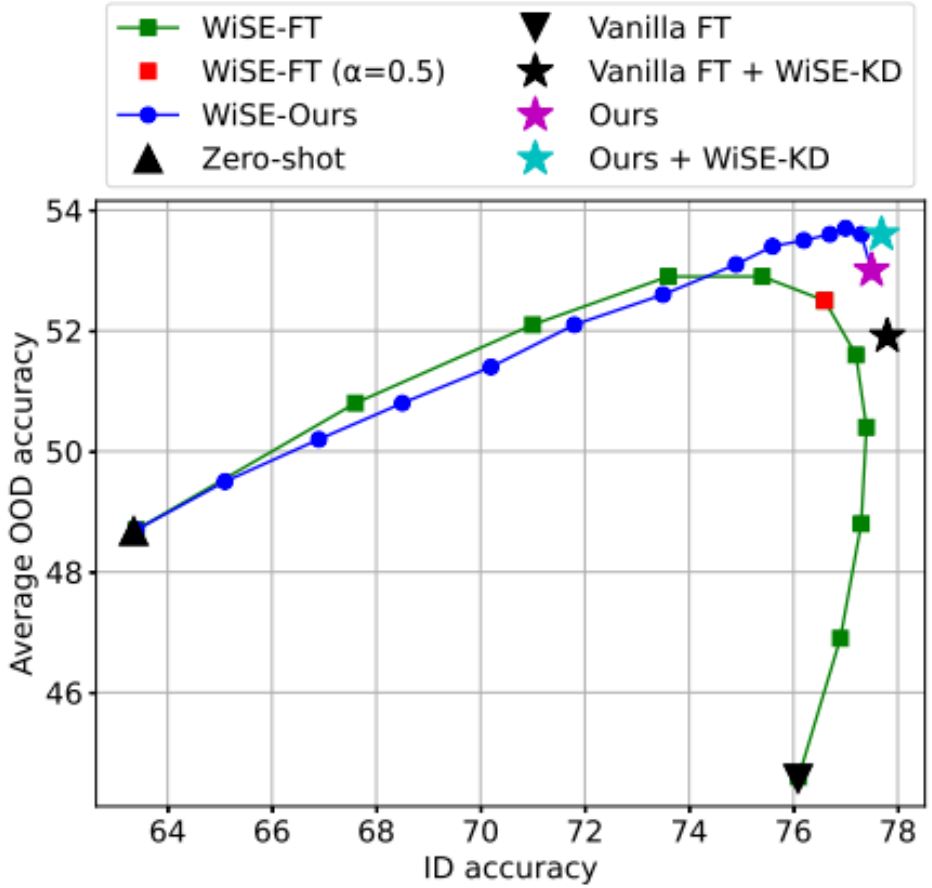
(Image source: [1])

[1] Wortsman, Mitchell, et al. "Robust fine-tuning of zero-shot models." CVPR, 2022.

Discussion: WiSE-FT [1]



(Image source: [1])



- Weight-space ensemble of the zero-shot model & our model is less meaningful
- Adding an distillation loss with WiSE-FT teacher slightly improves our model

[1] Wortsman, Mitchell, et al. "Robust fine-tuning of zero-shot models." CVPR, 2022.

Conclusion

- The **spurious correlation** between semantic and non-semantic factors in downstream data may account for the **robustness degradation** in fine-tuning.

- **Masked images** can be effective **counterfactual samples** for robust fine-tuning, breaking the spurious correlation.

- **Weight-space constraints** may be sufficient but **not necessary** for maintaining the robustness of the pre-trained model.

Thanks!

Paper: https://arxiv.org/abs/2303.03052

Project/Code: https://github.com/Coxy7/robust-finetuning/