ForAug: Recombining Foregrounds and Backgrounds to Improve Vision Transformer Training with Bias Mitigation - Supplementary Material -

Anonymous ICCV submission

Paper ID 6426 - Supplementary

Abstract

This is the supplementary material for the paper: ForAug: Recombining Foregrounds and Backgrounds to Improve Vision Transformer Training with Bias Mitigation

A. Training Setup

On ImageNet we use the same training setup as [1] and [2] without pretraining. As our focus is on evaluating the changes in accuracy due to *ForAug*/*ForNet*, like [1], we stick to one set of hyperparameters for all models. We list the settings used for training on ImageNet and *ForNet* in Table 1 and the ones used for finetuning those weights on the downstream datasets in Table 2.

Parameter	Value
Image Resolution	224×224
Epochs	300
Learning Rate	3e-3
Learning Rate Schedule	cosine decay
Batch Size	2048
Warmup Schedule	linear
Warmup Epochs	3
Weight Decay	0.02
Label Smoothing	0.1
Optimizer	Lamb [3]
Data Augmentation Policy	3-Augment [2]

Table 1. Training setup for our ImageNet and *ForNet* training.

Dataset	Batch Size	Epochs	Learning Rate
Aircraft	512	500	3e-4
Cars	1024	500	3e-4
Flowers	256	500	3e-4
Food	2048	100	3e-4
Pets	512	500	3e-4

Table 2. Training setup for finetuning on different downstream datasets. Other settings are the same as in Table 1.

B. Infill Model Comparison

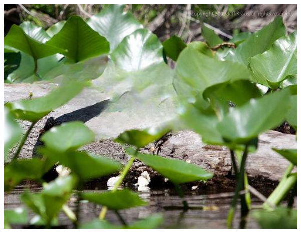
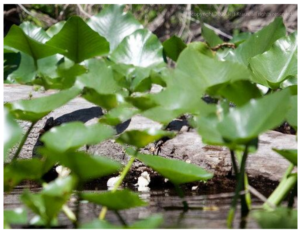
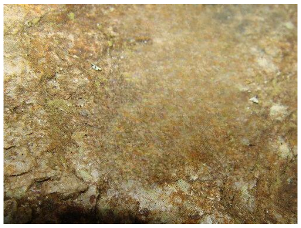
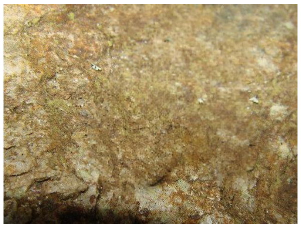
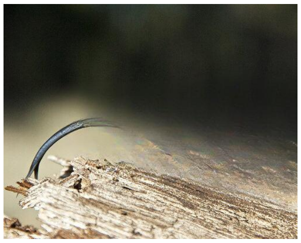
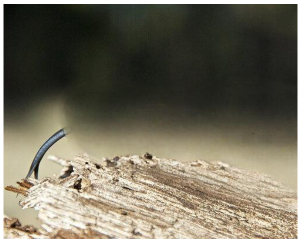
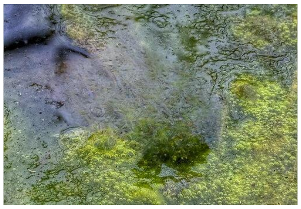
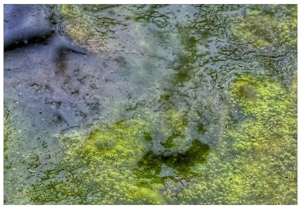
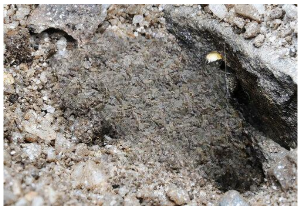
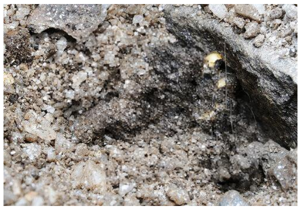
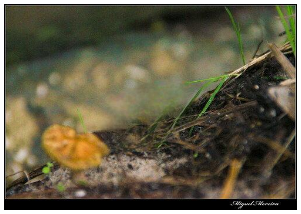
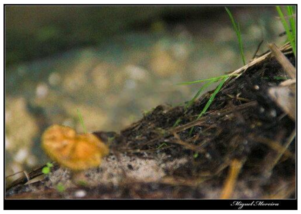
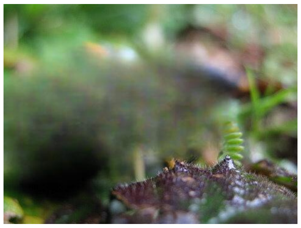
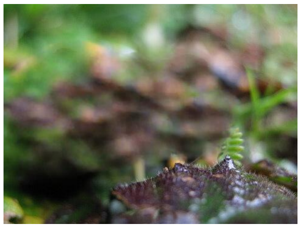
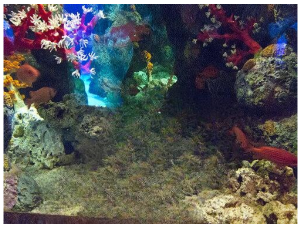
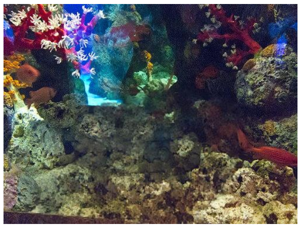
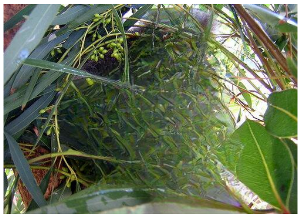
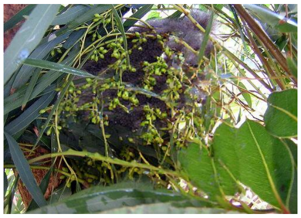
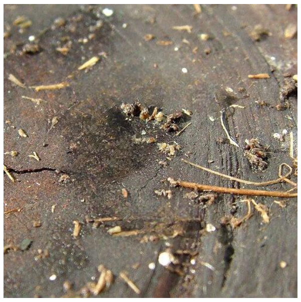
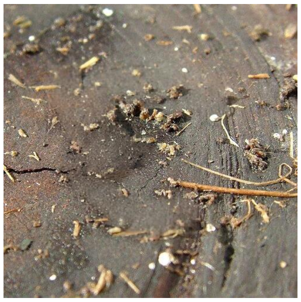
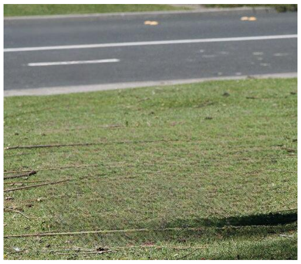
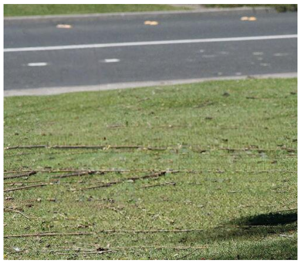
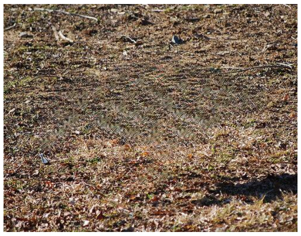
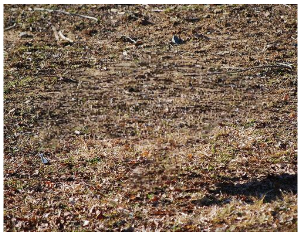
LaMa	Att. Eraser	LaMa	Att. Eraser
			
			
			
			
			
			

Table 3. Example infills of LaMa and Attentive Eraser.

C. Images with High Infill Ratio

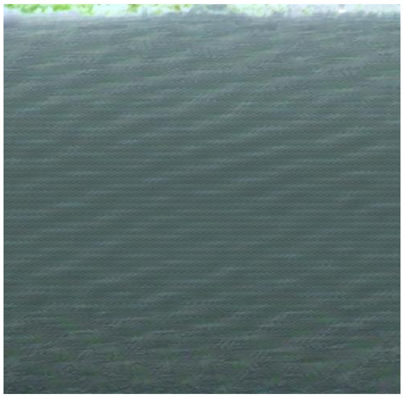
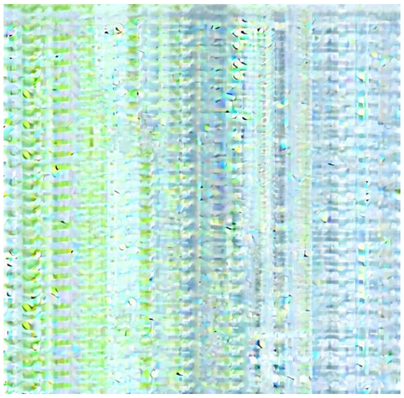
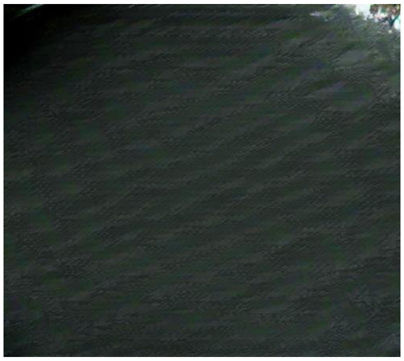
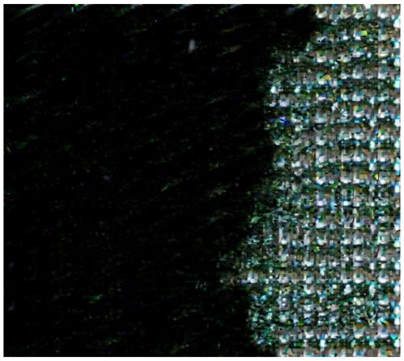
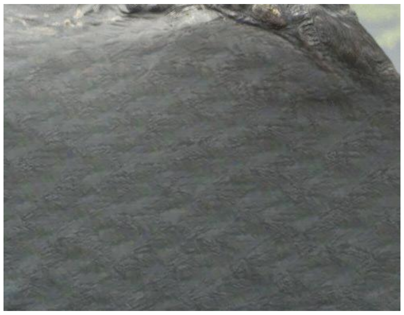
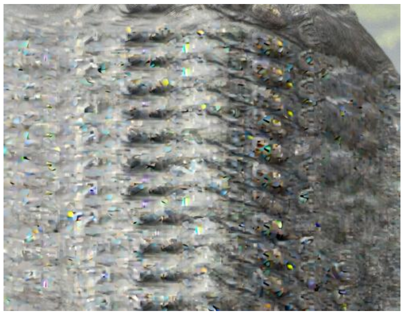
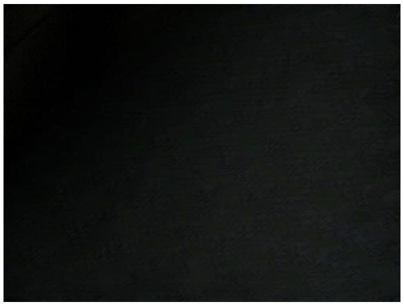
Infill Ratio	LaMa	Att. Eraser
93.7		
95.7		
83.7		
88.2		

Table 4. Example infills with a large relative foreground area size that is infilled (infill ratio).

References

[1] Tobias Christian Nauen, Sebastian Palacio, and Andreas Dengel. Which transformer to favor: A comparative analysis of efficiency in vision transformers, 2023. 1

[2] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *Computer Vision – ECCV 2022*, pages 516–533, Cham, 2022. Springer Nature Switzerland. 1

[3] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*, 2020. 1