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

We would like to sincerely thank the reviewers (**kCub**, **W3SS**, **5E96**) for their time and valuable feedback. Below, we will address each of the reviewers points.

Reasoning and purpose of *ForAug* (kCub): Traditional data augmentations are limited by existing image compositions, leading to biases where objects are centered and correlated with specific backgrounds. *ForAug* aims to overcome these limitations by introducing object size, position, and background as independent, controllable degrees of freedom. This approach explicitly exposes the model to a wider range of variations, actively reducing such compositional biases (see Tbls. 6, 8; Figs. 4, 1 (right)). Consequently, models trained with *ForAug* exhibit better performance specifically on these lower-likelihood images. Moreover, *ForAug* serves as an analytical tool for measuring biases in any ImageNet-trained model (Sec. 4.3). Acknowledging the need for more clarity, we have expanded Sec. 1 and 3 to further highlight the purpose and design, and revised Sec. 4.3 to more clearly connect experimental findings to *ForAug*'s goals.

Novelty of ForAug (kCub, W3SS): While inspired by Copy-Paste methods, *ForAug* makes distinct contributions by addressing the non-trivial challenges of classification, where we are successfully *"automating the copy-paste augmentation with [...] solid empirical gains"* (**5E96**). We want to emphasize some elements that make *ForAug* novel: **(1)** Adapting copy-paste to image classification has only been tried by [1] as an alternative to MixUp in a specialized domain. The scarcity of such methods suggests either *ForAug*'s novelty, or the inherent difficulty in achieving successful application, thereby highlighting the novelty of our specific design choices. **(2)** We overcome key challenges in adapting to classification. For label integrity we generate plain background images, removing the main object. Pasting a new object onto these backgrounds allows for a clear, unambiguous label. This approach, unlike [1] and previous Copy-Paste methods pasting onto existing dataset images, ensures clear training signals and reduces spurious background correlations. **(3)** *ForAug* incorporates large-scale position and size augmentations for the foregrounds to encode these equivariances into the training data for bias-mitigation, a feature not utilized to the same extent in [1] or [11, 14, 55].

Directly comparing to Copy-Paste (W3SS): Adapting Copy-Paste for classification brings several challenges, due to **(1)** its dependence on human-annotated foreground masks [14,28,41,53], which are generally not available for large-scale datasets used in image classification. **(2)** the difficulty in determining the augmented image's label. Pasting new foregrounds on existing dataset images creates label ambiguity (e.g., should the label derive from the new object, original, a multi-label?), unlike in segmentation where instance labels are preserved. Thus, "directly" applying Copy-Paste requires many design choices, essentially leading to a novel method.

[1] J. -S. Kang and K. Chung, "STAug: Copy-Paste Based Image Augmentation Technique Using Salient Target," in IEEE Access, vol. 10, 2022

Model	DeiT	Ours (DeiT)		Ours (DeiT III)	
	original	IN	FN	IN	FN
ViT-S	79.8*	80.5	80.3	79.1	**81.4**
ViT-B	81.8*	79.6	**81.5**	77.6	81.1
Swin-S	-	82.2	**82.4**	79.4	80.6

Table 1. Results when training on ImageNet (IN) and *ForNet* (FN) using different data augmentation schemes. DeiT uses EMA*.

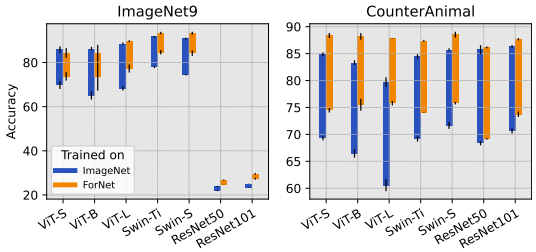

Figure 1. Results on ImageNet9 and CounterAnimal. Bars span from out-of-distribution (OOD, bottom) to normal backgrounds (top). *ForNet* (orange) significantly improves OOD performance compared to ImageNet (blue), reducing the accuracy gap (bar size).

Using different training pipelines (W3SS): While *ForAug* improves accuracy, our data augmentation pipeline does indeed not reach the results from DeiT. However, there is no reason to believe that *ForAug* does not work or improve the model performance when using another data augmentation or that this does make the comparison unfair, since the only difference between the results on ImageNet and *ForNet* is the inclusion of *ForAug*. Tab. 1 (above) presents results for a subset of models (time constraints) using the DeiT-pipeline, finding that *ForAug* still improves performance, especially for larger transformers, with the added benefit of bias reduction. Full results will be added to the manuscript.

Background Robustness using other datasets (5E96): While our metric (Eq. 4) was designed to mitigate *ForNet*-bias by comparing relative accuracy drops using the same recombination scheme, we agree and added the suggested benchmarks to the final manuscript (see Fig. 1 above). These new results support our findings, as *ForNet* reduces the background gap of transformers by boosting OOD performance.

Additional compute and space costs (5E96): We added a discussion to the manuscript. With ViT-B/16 on ImageNet (A100), *ForAug* leads to a minor 1% increase in average step-time (528 ± 2 ms to 534 ± 1 ms) since the online recombination is CPU-outsourced and heavily parallelized. *ForNet* requires 73GB of disk space, while ImageNet needs 147GB.