We thank the reviewers for their constructive feedback. We are pleased that they appreciated our "novel paper that is well motivated and executed" [R4], that "tackles an important and challenging problem of few-shot fine-grained classification" [R3], and that "will draw impact to both rigorous communities of few-shot and fine-grain recognition" [R4]. "The experiments, as well as the pilot study, are in great shape" [R4]. Our "framework works well on reasonably difficult datasets" [R1], and "can be useful for the future research" [R1]. Additionally, we would like to highlight a key contribution of our work: while GAN-generated images have not generally been useful for training image recognition models [26], we show how to effectively use such generated images for one-shot learning for fine-grained recognition.

R1: Class label input of BigGAN: We optimize the class conditional embedding and regard it as part of the input noise. Generally speaking, a conditional GAN uses input noise conditioned on the label of the image to generate. BigGAN also follows this approach, but our fine-tuning technique uses a single image to train. In other words, we only have a single class label and can then optimize the class embedding as part of the input noise. We will clarify this point.

R1: Compare with Mixup: As suggested, we ran Mixup as a baseline and obtained one-shot accuracy of 82.65 ± 0.59 and 88.12 ± 0.52 on CUB and NAB, respectively. These results are higher than baselines but still lower than ours.

R3: Compare with [7]: Data augmentation is also used in [7], published at CVPR 2019. The key difference is that while they augment support image by fusing with external real images from a gallery set, our model fuses with images synthesized by GANs, and it is non-trivial to make this change. Further, to adapt a generic pretrained GAN to a new domain, we introduce a technique of optimizing only z and BatchNorm parameters rather than the full generator, which is a novel aspect compared to [7]. Nevertheless, as suggested, we implemented their approach and obtained accuracies of 82.84 ± 0.62 on CUB and 88.42 ± 0.59 on NAB, which is higher than the baselines but not as high as ours.

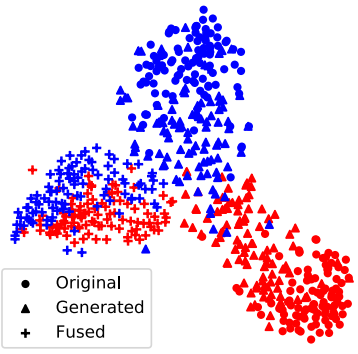
R3: Train end-to-end from scratch: Theoretically, we can do end-to-end training of all components, but in practice we are limited by our GPU memory, which is not large enough to hold both our model and BigGAN. However, to validate this point, we have added another experiment. We simplified BigGAN with one-quarter the number of channels and applied the rest of our approach with a backbone of four-layer CNN with random initialization. This model was trained end-to-end on CUB meta-training dataset, and achieved an accuracy of 63.77 ± 0.71, which is still higher than the baselines.

R3: Performance w.r.t. the number of synthesized images: As suggested, we increased $n_{aug} = 1, 2, 3, 5, 10$ on CUB and achieved accuracies 83.51 ± 0.60, 83.65 ± 0.60, 81.79 ± 0.62, 80.79 ± 0.62, 80.34 ± 0.63, and 79.75 ± 0.69, respectively. Too many augmented images seem to bias the classifier. We conclude that the performance gain is marginal or even harmful when increasing n_{aug}.

R3: Why our method works: We do not have firm theoretical explanations for why our method works beyond empirical evidence, but we would like to share some insights that guide us. The data distribution of GAN-generated images is biased towards frequent patterns (or modes) of the original image distribution [26], and may not help train image recognition. Our model helps diversify the data distribution by injecting artifacts of 3×3 patches, and thus can potentially help recognition. This point is supported by the t-SNE visualization below. R3 is correct that the fused images do not look very different from the originals for humans, but this might not be the case for CNNs; for example, adversarial noise is typically imperceptible to humans but dramatically changes CNN representations.

R4: More visualization and studies on learned augmentation: We will add more visualizations similar to Figure 3 in the supplemental material. We are happy to perform other suggestions for additional experiments.

R4: Fused image on feature space: As plotted in the right, we randomly pick two classes shown in red and blue, sample 100 images for each class, and apply t-SNE visualization of real images (•), generated images (▲), and augmented fused images (+). It is reasonable that the generated images are closer to the real ones, because our loss function (equation 1) encourages this to be so. Interestingly, perhaps due to artifacts of 3×3 patches, the fused images are distinctive from the real/generated images in the embedding space, extending the decision boundary.



R3, R4: Evaluation on ImageNet: Thank you for the suggestion, but we did not have enough computation resources to do this within the author response time period. Our goal was fine-grained recognition, which is why we did not perform ImageNet experiment originally. Nevertheless, we will include results on ImageNet in the camera ready. The public BigGAN model was trained on images including the meta-testing set of ImageNet, so we will have to train the model from scratch using meta-training set only.

Others: Thanks for the suggestions and corrections. We use one-vs-all classifier for logistic regression and will clarify.