We would like to thank all the reviewers for providing valuable feedback. Below are our responses to the comments.

Reviewer#1: 1) To the comment "the transfer scenarios in Sec 3 are confusing", we would like to explain that VGG-19 was indeed always used as the source model in Sec 3 in our paper. On lines 128–130, we meant to say that if we had trained the LinS model from scratch, the success rate of using it to attack VGG-19 (as shown in the grey curve in Figure 2) would have been much lower than fine-tuning (as shown in the grey curve in Figure 1). We will change the legends in Figure 1 and 2 to "VGG-19 $\rightarrow$ WRN", "VGG-19-LinS $\rightarrow$ WRN", etc. as suggested.

2) We followed the suggestion of performing experiments with ϵ=16/255, 8/255, and 4/255. On ImageNet, our LinBP achieved an average success rate of 84.77%, 50.56%, and 20.89%, respectively, showing that it still outperformed the other methods (e.g., ILA: 72.34%, 42.05%, and 17.60%) remarkably. In addition, when combined with ILA and SGM, our method further gained an average success rate of **90.20%** under ϵ=16/255. As has been recognized by the reviewer, we also reported results under ϵ=0.05 and 0.03 in our paper, and we will discuss the results further in these settings that lead to more imperceptible perturbations in the final version of the paper.

3) We followed the suggestion of testing with the momentum iterative FGSM (MI-FGSM) attack on both CIFAR-10 and ImageNet, and the superiority of our LinBP still held as with I-FGSM. Specifically, our LinBP achieved an average success rate of 87.50% on ImageNet while the second best method (i.e., ILA) achieved 71.21% in the untargeted setting under ϵ=16/255 (see Table 1b). We also reported the results using other baseline attacks (i.e., DI^2-FGSM, PGD, and an ensemble attack) in the supplementary material of the paper, which further demonstrate the effectiveness of our method.

4) We considered two other source models as suggested: ResNet-18 (on CIFAR-10) and Inception v3 (on ImageNet). With these two models, our method outperformed its competitors similarly under the constraint of ϵ=16/255, 8/255, and 4/255. See Table 1a and Table 1c for the detailed results.

5) We followed the suggestion of discussing targeted attacks. Table 1 shows that the superiority of our method holds on both CIFAR-10 and ImageNet in the targeted setting as well. Due to the space limit, we only compared our method with the baseline attack and the second best method in the table.

Table 1: More results of the transfer-based attacks on CIFAR-10 and ImageNet, using MI-FGSM as the baseline attack.

Source	Method	ϵ	Untargeted	Targeted
ResNet-18 (CIFAR-10)	MI-FGSM	16/255	84.35%	40.87%
		8/255	62.68%	28.84%
		4/255	34.00%	12.68%
	ILA	16/255	90.26%	39.19%
		8/255	73.75%	33.69%
		4/255	38.90%	14.49%
	LinBP (ours)	16/255	**94.03%**	**71.66%**
		8/255	**81.11%**	**57.24%**
		4/255	**47.32%**	**22.25%**

(a) CIFAR-10: ResNet-18 $\rightarrow$ victims

Source	Method	ϵ	Untargeted	Targeted
ResNet-50 (ImageNet)	MI-FGSM	16/255	58.67%	0.17%
		8/255	34.51%	0.06%
		4/255	16.94%	0.01%
	ILA	16/255	71.21%	0.34%
		8/255	40.84%	0.07%
		4/255	17.86%	0.02%
	LinBP (ours)	16/255	**87.50%**	**5.01%**
		8/255	**55.87%**	**0.93%**
		4/255	**25.16%**	**0.06%**

(b) ImageNet: ResNet-50 $\rightarrow$ victims

Source	Method	ϵ	Untargeted	Targeted
Inception v3 (ImageNet)	MI-FGSM	16/255	48.44%	0.15%
		8/255	31.16%	0.06%
		4/255	17.00%	0.01%
	ILA	16/255	75.04%	0.25%
		8/255	46.78%	0.14%
		4/255	21.95%	0.02%
	LinBP (ours)	16/255	**81.07%**	**0.35%**
		8/255	**48.26%**	**0.17%**
		4/255	**22.56%**	**0.11%**

(c) ImageNet: Inception v3 $\rightarrow$ victims

6) We ran attacks for 100 iterations to ensure that all the methods achieved their best performance. The success rate of the methods decreased 5%-20% if we ran only 10 iterations. Indeed, sometimes our LinBP achieved slightly higher attack success rates in attacking the source models than those of I-FGSM, similar to an observation made in the ILA paper. This is likely because analytical gradients cannot represent nonlinear functional changes of f (caused by each perturbation step, which is as large as 1/255), as commented by Reviewer#2.

Reviewer#2: We will discuss the mentioned related work. Thanks for the reference.

Reviewer#3: 1) We followed the suggestion of attacking a black-box robust ResNet (https://bit.ly/2C9FJVM) guarded by PGD adversarial training. The experiment shows similarity superiority of attack using LinBP (victim error rate: 48.60%, 39.92%, and 37.10% with ϵ=0.1, 0.05, 0.03) to ILA (42.30%, 37.82%, and 36.60%). The model in Sec 5.2 guarded by ensemble adversarial training was obtained on GitHub (https://bit.ly/2XKfrkz), provided by Kurakin et al. 2) Without re-normalization, the performance of our method degraded to 81.36% (from 96.89%), under ϵ=0.1. The norm of gradient became much larger in the main stream of the residual network with $W_i W_{i+1}$ being calculated instead of $W_i M_i W_{i+1}$, so that the gradient flowing through the main stream dominated, which is undesirable according to SGM. 3) We followed the policy of fine-tuning in a PyTorch tutorial, and more details will be included in an updated version of the paper. 4) With a two step policy, our LinBP indeed achieved higher success rates (90.49%, 74.44%, and 52.95%) than those of ILA in Table 6.

Reviewer#4: 1) We compared different methods on VGG-19/ResNet-18 on CIFAR-10 and ResNet-50/Inception v3 on ImageNet (see Table 1 in this response). It can be seen that the superiority of our method holds on all these concerned architectures. 2) Our method was invoked at the same positions as for ILA for fairness. Our paper discussed how the performance of our method and LinBP varied with the choice of positions in Figure 4.