A Fine-tuning $f'$ with Frozen $h$

Empirical results in Section 3.1 in the main paper show that simply removing ReLUs lead to improved transferability. In this section, we try freezing all learnable parameters in the unmodified sub-net $h$ during fine-tuning and a similar observation about the initial improvement of transferability can still be made (see Figure 5). Yet, we see obviously in Figure 5 that the curves bear a longer period of decrease and finally the obtained success rates are even worse than those of the baseline, owing to its lower learning capacity if $h$ is frozen. Classification loss of these modified VGG-19 models on the benign CIFAR-10 test set is also reported, in Figure 6.

Figure 5: How the transferability of (a) FGSM and (b) I-FGSM adversarial examples changes while fine-tuning $g'$ (i.e., the purely linear blocks of $f'$). $h$ is frozen during the fine-tuning process. We perform untargeted attacks under the constraint of $\ell_\infty$ constraints and $\epsilon$ is set as 0.03.

Figure 6: Test classification loss of different source models tested here and in Section 3.1.

*The first two authors contributed equally to this work, and Yiwen Guo is the corresponding author.
### B Benign-test-set Performance of The Victim Models

Table 5: Benign-test-set accuracy of the models. On ImageNet, it is evaluated on the 50,000 official validation images.

| Dataset    | VGG-19 (2015) | WRN (2016) | ResNeXt (2017) | DenseNet (2017) | PyramidNet† (2019) | GDAS (2019) |
|------------|---------------|------------|----------------|------------------|--------------------|-------------|
| CIFAR-10   | 93.34%        | 96.21%     | 96.31%         | 96.68%           | 98.44%             | 97.19%      |
| ImageNet   | 76.15%        | 77.45%     | 74.65%         | 71.88%           | 82.90%             | 81.32%      |

### C Re-interpret Intermediate Feature-based Attacks?

As mentioned in the main paper, many recent successes in improving adversarial transferability benefit from maximizing intermediate level distortions rather than the final prediction losses of DNNs. Taking ILA [2] as an example, it opts to maximizing the projection of disturbance \( h(x + r) - h(x) \) on a guidance direction \( \mathbf{v} := h(x + r_0) - h(x) \), given a prior adversarial perturbation \( r_0 \), i.e.,

\[
\max_{\|\mathbf{r}\| \leq \varepsilon} J(x + r, y) \text{ s.t. } J(x + r, y) = \mathbf{v}^T(h(x + r) - h(x)).
\]  

Recall that many adversarial attacks attempt to solve the optimization problem (1). With the common decomposed representation \( f = g \circ h \), we can write \( L = l \circ s \circ g \circ h \), in which \( s \) is the softmax function and \( l \) is the cross-entropy loss that takes both the output of \( s \) and the one-hot encoding of \( y \) as inputs. Apparently, \( J \) in Eq. (4) is more linear than \( L \) on account of the additional non-linearity of \( l, s, \) and \( g \). It can be considered as substituting \( s \circ g \) with an identity mapping and replacing \( l \) with a linear loss. Likewise, AA [3] also exploits an identity mapping, in combination with a quadratic loss though. Whether their practical effectiveness partially comes from higher linearity worth exploring in future work.

### D Single-step Back-end Attacks

When a single-step back-end attack FGSM is adopted, different methods are compared in Table 6 and 7. It can be seen that the superiority of method holds in most cases as with I-FGSM. It is worthy of mentioning that ILA is performed after the FGSM examples are obtained in advance, and thus it takes one more step in comparison to LinBP and others. We verified this by using a two-step policy for our LinBP on CIFAR-10, and it showed that our LinBP indeed achieved higher average success rates (90.49%, 74.44%, and 52.95%) than those of ILA in Table 6 (85.72%, 70.94%, and 50.68%).

Table 6: Success rates of transfer-based attacks on CIFAR-10 using FGSM with \( \ell_\infty \) constraints in the untargeted setting. TAP boils down to FGSM in the single-step setting hence their performance are the same. For ILA, we test it on the basis of FGSM examples crafted offline and take one more step following its optimization objective, i.e., \( \text{two steps in total} \). The source model is a VGG-19 with batch normalization and the symbol * indicates that the victim model is the same as the source model. Average is obtained from models different from the source.

| Dataset    | Method       | \( \epsilon \) | VGG-19* (2015) | WRN (2016) | ResNeXt (2017) | DenseNet (2017) | PyramidNet† (2019) | GDAS (2019) | Average |
|------------|--------------|----------------|----------------|------------|----------------|------------------|--------------------|-------------|---------|
| CIFAR-10   | FGSM         | 0.1            | 82.52%         | 74.98%     | 81.64%         | 84.28%           | 61.34%             | 88.86%      | 78.22% |
|            |              | 0.05           | 73.38%         | 55.34%     | 61.64%         | 59.00%           | 32.50%             | 67.18%      | 55.13% |
|            |              | 0.03           | 66.68%         | 39.52%     | 42.36%         | 39.86%           | 17.42%             | 39.00%      | 35.63% |
|            | TAP+FGSM     | 0.1            | 82.52%         | 74.98%     | 81.64%         | 84.28%           | 61.34%             | 88.86%      | 78.22% |
|            |              | 0.05           | 73.38%         | 55.34%     | 61.64%         | 59.00%           | 32.50%             | 67.18%      | 55.13% |
|            |              | 0.03           | 66.68%         | 39.52%     | 42.36%         | 39.86%           | 17.42%             | 39.00%      | 35.63% |
|            | ILA+FGSM     | 0.1            | 88.92%         | 85.32%     | 87.46%         | 89.30%           | 76.68%             | 89.82%      | 85.72% |
|            |              | 0.05           | 87.14%         | 73.9%      | 76.48%         | 76.82%           | 48.12%             | 79.40%      | 70.94% |
|            |              | 0.03           | 82.04%         | 55.99%     | 59.64%         | 58.36%           | 24.76%             | 54.66%      | 50.68% |
|            | LinBP+FGSM   | 0.1            | 92.06%         | 84.16%     | 87.24%         | 87.76%           | 72.36%             | 89.38%      | 84.18% |
|            |              | 0.05           | 95.18%         | 73.96%     | 77.60%         | 76.34%           | 43.76%             | 77.50%      | 69.83% |
|            |              | 0.03           | 92.90%         | 55.64%     | 59.70%         | 55.82%           | 22.30%             | 52.28%      | 49.15% |
We notice that there exist different setting in evaluations of adversarial attacks in recent papers, some with them than with I-FGSM. See Table 8 and 9 for detailed results when combining our LinBP with ImageNet.

| Dataset | Method | ResNet* (2016) | Inception v3 (2016) | DenseNet (2017) | MobileNet v2 (2018) | PNASNet (2018) | SENet (2018) | Average |
|---------|--------|----------------|---------------------|-----------------|---------------------|----------------|-------------|---------|
| FGSM    | 0.1    | 88.64%         | 47.88%              | 63.86%          | 79.80%              | 93.03%         | 48.96%      | 55.96%  |
|         | 0.05   | 86.46%         | 32.52%              | 44.96%          | 50.34%              | 24.36%         | 28.26%      | 36.09%  |
|         | 0.03   | 84.38%         | 23.04%              | 35.26%          | 37.72%              | 16.44%         | 18.74%      | 26.24%  |
| TAP+FGSM| 0.1    | 88.64%         | 47.88%              | 63.86%          | 79.80%              | 93.03%         | 48.96%      | 55.96%  |
|         | 0.05   | 86.46%         | 32.52%              | 44.96%          | 50.34%              | 24.36%         | 28.26%      | 36.09%  |
|         | 0.03   | 84.38%         | 23.04%              | 35.26%          | 37.72%              | 16.44%         | 18.74%      | 26.24%  |
| ILA+FGSM| 0.1    | 84.38%         | 50.40%              | 67.10%          | 82.12%              | 93.03%         | 51.60%      | 58.86%  |
|         | 0.05   | 74.56%         | 30.56%              | 44.14%          | 55.44%              | 19.99%         | 27.28%      | 35.48%  |
|         | 0.03   | 74.38%         | 20.68%              | 32.64%          | 38.66%              | 13.26%         | 16.54%      | 24.36%  |
| SGM+FGSM| 0.1    | 84.30%         | 28.04%              | 40.90%          | 50.74%              | 19.88%         | 25.68%      | 33.05%  |
|         | 0.05   | 82.96%         | 19.72%              | 31.74%          | 37.72%              | 12.70%         | 16.42%      | 23.66%  |
|         | 0.03   | 82.64%         | 10.68%              | 22.64%          | 38.66%              | 13.26%         | 16.54%      | 24.36%  |
| LinBP+FGSM| 1.0   | 91.42%         | 52.64%              | 69.44%          | 83.98%              | 92.66%         | 52.72%      | 60.34%  |
|         | 0.05   | 90.52%         | 34.80%              | 49.24%          | 56.26%              | 25.78%         | 31.18%      | 39.45%  |
|         | 0.03   | 88.56%         | 24.80%              | 39.28%          | 41.68%              | 17.32%         | 19.94%      | 28.60%  |

Table 7: Success rates of transfer-based attacks on ImageNet using FGSM with \( \ell_{\infty} \) constraints. TAP boils down to FGSM in the single-step setting thus their performance are the same. For ILA, we test it on the basis of FGSM examples crafted in advance and take one more step following its optimization objective, i.e., two steps in total. The source model is a ResNet-50 and the symbol * indicates that the victim model is the same as the source. Average is obtained from models different from the source.

Table 8: Success rates of transfer-based attacks on ImageNet using DI\(^2\)-FGSM. The source model is a ResNet-50 and the symbol * indicates that the victim model is the same as the source model. The mean and standard deviation results of five runs are reported. Average is obtained from models different from the source.

Table 9: Success rates of transfer-based attacks on ImageNet using PGD. The source model is a ResNet-50 and the symbol * indicates that the victim model is the same as the source model. The mean and standard deviation results of five runs are reported. Average is obtained from models different from the source.

F  Other \( \epsilon \) Settings and More Source Models

We notice that there exist different setting in evaluations of adversarial attacks in recent papers, some use \( \epsilon = 0.1, 0.08, 0.05, 0.04 \), and 0.03 as in the main body of our paper [2, 5, 4], and some others use \( \epsilon = 16/255, 8/255, \) and 4/255, etc. Here we also report the performance of our
method under $\epsilon = 16/255$, $8/255$, and $4/255$ for easier comparison with contemporary methods and other comparable methods. We also considered more source architectures for comparison, including ResNet-18 on CIFAR-10 and Inception v3 on ImageNet. The results are shown in Table 10, Table 11 Table 12, and Table 13, in which PGD is used as the baseline, since it is more powerful than I-FGSM, and other methods were performed on the basis of PGD, just like in Table 9. Here we give results in both the untargeted and the targeted settings. Results on the basis of other back-end attacks are similar, and the superiority of our LinBP holds in most test cases.

**Table 10:** Success rates of transfer-based attacks on CIFAR-10 using PGD with $\ell_\infty$ constraints. The source model is a VGG-19. Average is obtained from models different from the source.

| Attack Method | $\epsilon$ | VGG-19* (2015) | WRN (2016) | ResNetXt (2017) | DenseNet (2017) | PyramidNet (2019) | GDAS (2019) | Average |
|---------------|------------|----------------|-------------|----------------|-----------------|------------------|-------------|---------|
| Targeted      |            |                |             |                |                 |                  |             |         |
| LinBP+PGD     | 16/255     | 99.98%         | 99.99%      | 99.56%         | 99.66%          | 99.18%           | 80.54%      | 94.92%  |
| (VGG-19)      | 8/255      | 99.92%         | 96.12%      | 99.94%         | 99.47%          | 99.14%           | 76.04%      | 94.75%  |
| LinBP         | 4/255      | 99.42%         | 96.12%      | 99.94%         | 98.96%          | 98.96%           | 71.64%      | 93.78%  |
| Untargeted    |            |                |             |                |                 |                  |             |         |
| LinBP+PGD     | 16/255     | 100.00%        | 100.00%     | 99.99%         | 99.99%          | 99.96%           | 99.54%      | 99.44%  |
| (VGG-19)      | 8/255      | 100.00%        | 99.56%      | 99.56%         | 99.94%          | 99.94%           | 94.86%      | 99.94%  |
| LinBP         | 4/255      | 99.82%         | 96.12%      | 99.94%         | 99.94%          | 99.94%           | 84.62%      | 97.88%  |

**Table 11:** Success rates of transfer-based attacks on CIFAR-10 using PGD with $\ell_\infty$ constraints. The source model is a ResNet-18. Average is obtained from models different from the source.

| Attack Method | $\epsilon$ | VGG-19 (2015) | WRN (2016) | ResNeXt (2017) | DenseNet (2017) | PyramidNet (2019) | GDAS (2019) | Average |
|---------------|------------|----------------|-------------|----------------|-----------------|------------------|-------------|---------|
| Targeted      |            |                |             |                |                 |                  |             |         |
| LinBP+PGD     | 16/255     | 94.20%         | 98.08%      | 97.76%         | 95.88%          | 72.82%           | 91.20%      | 91.66%  |
| (ResNet-18)   | 8/255      | 97.90%         | 96.62%      | 96.08%         | 96.16%          | 70.36%           | 81.70%      | 88.18%  |
| LinBP         | 4/255      | 97.46%         | 98.30%      | 97.40%         | 97.52%          | 74.20%           | 77.52%      | 83.53%  |
| Untargeted    |            |                |             |                |                 |                  |             |         |
| LinBP+PGD     | 16/255     | 94.54%         | 96.54%      | 95.42%         | 93.98%          | 78.48%           | 93.10%      | 92.01%  |
| (ResNet-18)   | 8/255      | 76.24%         | 89.76%      | 89.24%         | 87.28%          | 46.48%           | 76.92%      | 77.94%  |
| LinBP         | 4/255      | 37.74%         | 57.14%      | 62.66%         | 57.04%          | 15.18%           | 49.10%      | 45.64%  |

**Table 12:** Success rates of transfer-based attacks on ImageNet using PGD with $\ell_\infty$ constraints. The source model is a ResNet-50. Average is obtained from models different from the source.

| Attack Method | $\epsilon$ | ResNet* (2016) | Inception v3 (2016) | DenseNet (2017) | MobileNet v2 (2018) | PNASNet (2018) | SENet (2018) | Average |
|---------------|------------|----------------|----------------------|-----------------|---------------------|----------------|--------------|---------|
| Targeted      |            |                |                      |                 |                     |                |              |         |
| LinBP+PGD     | 16/255     | 100.00%        | 98.46%               | 75.22%          | 74.18%              | 46.34%         | 53.22%       | 59.46%  |
| (ResNet-50)   | 8/255      | 100.00%        | 81.10%               | 63.86%          | 63.52%              | 21.60%         | 36.22%       | 41.56%  |
| LinBP         | 4/255      | 100.00%        | 81.10%               | 63.86%          | 63.52%              | 21.60%         | 36.22%       | 41.56%  |
| Untargeted    |            |                |                      |                 |                     |                |              |         |
| LinBP+PGD     | 16/255     | 100.00%        | 81.10%               | 63.86%          | 63.52%              | 21.60%         | 36.22%       | 41.56%  |
| (ResNet-50)   | 8/255      | 100.00%        | 81.10%               | 63.86%          | 63.52%              | 21.60%         | 36.22%       | 41.56%  |
| LinBP         | 4/255      | 100.00%        | 81.10%               | 63.86%          | 63.52%              | 21.60%         | 36.22%       | 41.56%  |
Table 13: Success rates of transfer-based attacks on *ImageNet* using PGD with $\ell_\infty$ constraints. The source model is an *Inception v3*. Average is obtained from models different from the source.

| Attack  | Method      | $\epsilon$ | ResNet(2016) | Inception v3* (2016) | DenseNet v2 (2017) | MobileNet v2 (2018) | PNASNet (2018) | SENet (2018) | Average |
|---------|-------------|-------------|--------------|---------------------|------------------|--------------------|----------------|-------------|---------|
|         |             |             |              |                     |                  |                    |                |             |         |
| Untargeted (Inception) | PGD 16/255 | 16/255       | 47.46%       | **100.00%**         | 48.22%           | 53.72%            | 34.66%         | 36.50%      | 44.14%  |
|         |             | 8/255        | 57.38%       | **100.00%**         | 53.72%           | 65.00%            | 53.72%         | 84.94%      | 99.80%  |
|         |             | 4/255        | 21.64%       | **99.96%**          | 48.22%           | 73.22%            | 34.66%         | 36.50%      | 34.66%  |
| ILA+PGD | PGD 16/255  | 16/255       | 84.94%       | **99.80%**          | 90.00%           | 89.28%            | 65.00%         | 74.70%      | 78.96%  |
|         |             | 8/255        | 55.42%       | **99.80%**          | 90.00%           | 64.12%            | 33.82%         | 42.16%      | 40.12%  |
|         |             | 4/255        | 25.96%       | **99.76%**          | 90.00%           | 24.08%            | 13.14%         | 22.96%      | 22.96%  |
| LinBP+PGD | PGD 16/255 | 16/255       | 90.82%       | **99.96%**          | 92.68%           | 89.28%            | 78.02%         | 85.82%      | 87.32%  |
|         |             | 8/255        | 56.22%       | **99.76%**          | 92.68%           | 55.46%            | 34.78%         | 45.44%      | 51.40%  |
|         |             | 4/255        | 22.08%       | **96.28%**          | 92.68%           | 22.92%            | 10.04%         | 14.64%      | 19.90%  |
| Targeted (Inception) | PGD 16/255  | 16/255       | 0.12%        | **99.98%**          | 0.20%            | 0.08%             | 0.16%          | 0.12%       | 0.14%   |
|         |             | 8/255        | 0.02%        | **99.96%**          | 0.04%            | 0.02%             | 0.06%          | 0.06%       | 0.04%   |
|         |             | 4/255        | 0.00%        | **98.42%**          | 0.02%            | 0.00%             | 0.00%          | 0.00%       | 0.00%   |
| ILA+PGD | PGD 16/255  | 8/255        | 0.52%        | **2.02%**           | 0.64%            | 0.18%             | 0.56%          | 0.26%       | 0.43%   |
|         |             | 8/155        | 0.22%        | **14.54%**          | 0.30%            | 0.22%             | 0.28%          | 0.14%       | 0.23%   |
|         |             | 4/255        | 0.06%        | **52.48%**          | 0.04%            | 0.06%             | 0.04%          | 0.06%       | 0.05%   |
| LinBP+PGD | PGD 16/255 | 16/255       | 0.34%        | **4.26%**           | 0.44%            | **0.26%**         | 0.32%          | 0.28%       | 0.33%   |
|         |             | 8/255        | 0.10%        | **6.08%**           | 0.12%            | 0.14%             | 0.06%          | 0.20%       | 0.12%   |
|         |             | 4/255        | 0.02%        | **4.04%**           | 0.00%            | 0.00%             | 0.00%          | 0.04%       | 0.01%   |

References

[1] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018.

[2] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *ICCV*, 2019.

[3] Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *CVPR*, 2019.

[4] Qizhang Li, Yiwen Guo, and Hao Chen. Practical no-box adversarial attacks against dnns. In *NeurIPS*, 2020.

[5] Qizhang Li, Yiwen Guo, and Hao Chen. Yet another intermediate level attack. In *ECCV*, 2020.

[6] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[8] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqiu Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *ECCV*, 2018.