# Supplementary Material for: Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks

Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu*
Dept. of Comp. Sci. and Tech., BNRist Center, State Key Lab for Intell. Tech. & Sys.,
Institute for AI, THBI Lab, Tsinghua University, Beijing, 100084, China
{dyp17, pty17}@mails.tsinghua.edu.cn, {suhangss, dcszj}@mail.tsinghua.edu.cn

We first show the results of the proposed translation-invariant attack method for white-box attacks and black-box attacks against normally trained models. We adopt the same settings for attacks. We also generate adversarial examples for Inception v3 (Inc-v3) [5], Inception v4 (Inc-v4), Inception ResNet v2 (IncRes-v2) [4], and ResNet v2-152 (Res-v2-152) [2], respectively, using FGSM, TI-FGSM, MI-FGSM, TI-MI-FGSM, DIM, and TI-DIM. For the translation-invariant based attacks, we use the $7 \times 7$ Gaussian kernel, since that the normally trained models have similar discriminative regions. We then use these adversarial examples to attack six normally trained models—Inc-v3, Inc-v4, IncRes-v2, Res-v2-152, VGG-16 [3], and Res-v1-152 [1]. The results are shown in Table 6 for FGSM and TI-FGSM, Table 7 for MI-FGSM and TI-MI-FGSM, and Table 8 for DIM and TI-DIM. The translation-invariant based attacks get better results in most cases than the baseline attacks.

Moreover, the experiments above and in the main paper are conducted based on the $L_\infty$ norm bound. We further demonstrate the applicability of the proposed method for other norm bounds, especially the $L_2$ norm bound. Similar to the results in Table 2-5, we present the results of FGSM and TI-FGSM in Table 9, MI-FGSM and TI-MI-FGSM in Table 10, DIM and TI-DIM in Table 11, and the ensemble method in Table 12. All those results are based on the $L_2$ norm bound, and we set the maximum perturbation $\epsilon = 10 \cdot \sqrt{d}$, where $d$ is the dimension of input images. The results based on the $L_2$ norm bound also show the effectiveness of the proposed method.

## References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 1

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[4] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. 1

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 1

*Corresponding author.

| | Attack | Inc-v3 | Inc-v4 | IncRes-v2 | Res-v2-152 | VGG-16 | Res-v1-152 |
|---|---|---|---|---|---|---|---|
| Inc-v3 | FGSM | **79.6*** | 35.9 | 30.6 | 30.2 | 49.7 | 36.3 |
| | TI-FGSM | 75.4* | **37.3** | **32.1** | **34.1** | **62.0** | **44.9** |
| Inc-v4 | FGSM | 43.1 | **72.6*** | 32.5 | 34.3 | 50.7 | 37.7 |
| | TI-FGSM | **45.3** | 68.1* | **33.7** | **35.4** | **63.3** | **46.2** |
| IncRes-v2 | FGSM | 44.3 | 36.1 | **64.3*** | 31.9 | 49.4 | 38.6 |
| | TI-FGSM | **49.7** | **41.5** | 63.7* | **40.1** | **64.2** | **46.7** |
| Res-v2-152 | FGSM | 40.1 | 34.0 | 30.3 | **81.3*** | 50.5 | 40.8 |
| | TI-FGSM | **46.4** | **39.3** | **33.4** | 78.9* | **64.7** | **50.4** |

Table 6. The success rates (%) of adversarial attacks against six normally trained models—Inc-v3, Inc-v4, IncRes-v2, Res-v2-152, VGG-16, and Res-v1-152. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2, and Res-v2-152, respectively, using FGSM and TI-FGSM. * indicates the white-box attacks.

| | Attack | Inc-v3 | Inc-v4 | IncRes-v2 | Res-v2-152 | VGG-16 | Res-v1-152 |
|---|---|---|---|---|---|---|---|
| Inc-v3 | MI-FGSM | 97.8* | 47.1 | 46.4 | 38.7 | 50.3 | 38.1 |
| | TI-MI-FGSM | **97.9*** | **52.4** | **47.9** | **41.1** | **63.4** | **48.1** |
| Inc-v4 | MI-FGSM | 67.1 | **98.8*** | 54.3 | 47.0 | 58.5 | 43.2 |
| | TI-MI-FGSM | **68.6** | **98.8*** | **55.3** | **47.7** | **69.0** | **51.3** |
| IncRes-v2 | MI-FGSM | 74.8 | 64.8 | **100.0*** | 54.5 | 59.3 | 50.8 |
| | TI-MI-FGSM | **76.1** | **69.5** | **100.0*** | **59.6** | **74.4** | **61.5** |
| Res-v2-152 | MI-FGSM | 54.2 | 48.1 | 44.3 | **97.5*** | 52.6 | 48.7 |
| | TI-MI-FGSM | **55.6** | **50.9** | **45.1** | 97.4* | **65.6** | **59.6** |

Table 7. The success rates (%) of adversarial attacks against six normally trained models—Inc-v3, Inc-v4, IncRes-v2, Res-v2-152, VGG-16, and Res-v1-152. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2, and Res-v2-152, respectively, using MI-FGSM and TI-MI-FGSM. * indicates the white-box attacks.

| | Attack | Inc-v3 | Inc-v4 | IncRes-v2 | Res-v2-152 | VGG-16 | Res-v1-152 |
|---|---|---|---|---|---|---|---|
| Inc-v3 | DIM | 98.3* | 73.8 | 67.8 | 58.4 | 62.5 | 49.3 |
| | TI-DIM | **98.5*** | **75.2** | **69.2** | **59.0** | **74.3** | **59.1** |
| Inc-v4 | DIM | **81.8** | 98.2* | **74.2** | **65.1** | 65.5 | 51.4 |
| | TI-DIM | 80.7 | **98.7*** | 73.2 | 62.7 | **77.4** | **59.8** |
| IncRes-v2 | DIM | 86.1 | 83.5 | **99.1*** | 73.5 | 67.9 | 62.7 |
| | TI-DIM | **86.4** | **85.5** | 98.8* | **76.3** | **79.3** | **72.2** |
| Res-v2-152 | DIM | **77.0** | **77.8** | **73.5** | **97.4*** | 67.4 | 67.8 |
| | TI-DIM | **77.0** | 73.9 | 73.2 | 97.2* | **78.4** | **77.8** |

Table 8. The success rates (%) of adversarial attacks against six normally trained models—Inc-v3, Inc-v4, IncRes-v2, Res-v2-152, VGG-16, and Res-v1-152. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2, and Res-v2-152, respectively, using DIM and TI-DIM. * indicates the white-box attacks.

| | Attack | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ | HGD | R&P | JPEG | TVM | NIPS-r3 |
|---|---|---|---|---|---|---|---|---|---|
| Inc-v3 | FGSM | 13.7 | 14.5 | 6.8 | 6.0 | 6.1 | 10.9 | 22.0 | 8.2 |
| | TI-FGSM | **15.2** | **15.7** | **10.2** | **8.2** | **18.8** | **11.0** | **25.7** | **10.4** |
| Inc-v4 | FGSM | **13.9** | 15.0 | 8.2 | **8.3** | 7.4 | **11.5** | 22.2 | 8.5 |
| | TI-FGSM | **13.9** | **16.2** | **10.4** | 8.0 | **9.1** | 11.3 | **24.3** | **8.9** |
| IncRes-v2 | FGSM | 16.0 | 17.5 | 11.3 | 10.8 | 10.2 | 14.4 | 26.2 | 11.6 |
| | TI-FGSM | **18.1** | **18.5** | **15.5** | **12.3** | **13.2** | **14.7** | **29.4** | **13.6** |
| Res-v2-152 | FGSM | 12.7 | 15.1 | 8.1 | 7.0 | 7.1 | 10.2 | 20.3 | 8.2 |
| | TI-FGSM | **13.4** | **15.8** | **9.7** | **7.2** | **7.9** | **10.7** | **22.5** | **9.1** |

Table 9. The success rates (%) of black-box attacks against eight defenses based on the $L_2$ norm bound. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2, and Res-v2-152 respectively using FGSM and TI-FGSM.

| | Attack | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ | HGD | R&P | JPEG | TVM | NIPS-r3 |
|---|---|---|---|---|---|---|---|---|---|
| Inc-v3 | MI-FGSM | 15.9 | 16.3 | 7.0 | 7.8 | 7.5 | 12.8 | 15.7 | 8.4 |
| | TI-MI-FGSM | **22.8** | **24.6** | **14.8** | **14.0** | **13.0** | **15.8** | **22.7** | **15.1** |
| Inc-v4 | MI-FGSM | 18.1 | 18.7 | 8.3 | 9.3 | 9.0 | 14.9 | 17.5 | 10.7 |
| | TI-MI-FGSM | **24.3** | **25.5** | **27.9** | **15.7** | **15.9** | **29.0** | **25.2** | **16.5** |
| IncRes-v2 | MI-FGSM | 22.9 | 21.6 | 16.6 | 17.1 | 15.2 | 22.2 | 20.9 | 18.0 |
| | TI-MI-FGSM | **35.0** | **35.8** | **30.5** | **26.3** | **26.4** | **29.8** | **35.6** | **28.8** |
| Res-v2-152 | MI-FGSM | 18.6 | 18.7 | 10.4 | 12.4 | 10.8 | 14.9 | 15.9 | 11.1 |
| | TI-MI-FGSM | **21.6** | **23.3** | **17.3** | **15.1** | **15.6** | **18.7** | **24.6** | **17.6** |

Table 10. The success rates (%) of black-box attacks against eight defenses based on the $L_2$ norm bound. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2, and Res-v2-152 respectively using MI-FGSM and TI-MI-FGSM.

| | Attack | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ | HGD | R&P | JPEG | TVM | NIPS-r3 |
|---|---|---|---|---|---|---|---|---|---|
| Inc-v3 | DIM | 17.9 | 21.8 | 9.7 | 11.8 | 10.0 | 15.5 | 17.0 | 12.7 |
| | TI-DIM | **29.6** | **31.9** | **22.0** | **20.1** | **20.0** | **22.0** | **27.3** | **23.9** |
| Inc-v4 | DIM | 21.6 | 22.2 | 12.9 | 15.8 | 13.3 | 20.5 | 19.2 | 16.6 |
| | TI-DIM | **31.0** | **33.1** | **24.0** | **22.8** | **22.9** | **24.8** | **29.2** | **25.1** |
| IncRes-v2 | DIM | 34.5 | 31.0 | 23.8 | 27.0 | 25.8 | 31.5 | 25.0 | 26.9 |
| | TI-DIM | **43.3** | **45.2** | **42.4** | **39.3** | **42.7** | **42.2** | **43.3** | **41.2** |
| Res-v2-152 | DIM | 29.0 | 30.1 | 18.7 | 27.8 | 19.8 | 26.7 | 21.3 | 23.1 |
| | TI-DIM | **36.3** | **37.2** | **28.9** | **28.0** | **30.0** | **28.4** | **36.1** | **32.7** |

Table 11. The success rates (%) of black-box attacks against eight defenses based on the $L_2$ norm bound. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2, and Res-v2-152 respectively using DIM and TI-DIM.

| Attack | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ | HGD | R&P | JPEG | TVM | NIPS-r3 |
|---|---|---|---|---|---|---|---|---|
| FGSM | **26.6** | **27.3** | 16.0 | **18.1** | 16.5 | **21.1** | 23.7 | 17.9 |
| TI-FGSM | 26.1 | 26.7 | **19.2** | 17.1 | **19.1** | 20.0 | **27.2** | **19.1** |
| MI-FGSM | 44.3 | 42.8 | 27.2 | 40.7 | 28.1 | 43.6 | 30.8 | 34.4 |
| TI-MI-FGSM | **59.3** | **59.0** | **53.0** | **54.6** | **50.0** | **53.3** | **51.3** | **51.1** |
| DIM | 57.0 | 54.7 | 37.4 | 58.9 | 43.4 | **60.3** | 37.3 | 50.3 |
| TI-DIM | **66.9** | **66.0** | **60.4** | **63.2** | **62.9** | 58.4 | **58.4** | **62.7** |

Table 12. The success rates (%) of black-box attacks against eight defenses based on the $L_2$ norm bound. The adversarial examples are crafted for the ensemble of Inc-v3, Inc-v4, IncRes-v2, and Res-v2-152 using FGSM, TI-FGSM, MI-FGSM, TI-MI-FGSM, DIM, and TI-DIM.