# Supplementary: Feature Importance-aware Transferable Adversarial Attacks

## Abstract

*To intuitively demonstrate the effectiveness of the proposed FIA, qualitative comparison is provided in this supplementary. Following the experimental evaluation in the main submission, corresponding examples in attacking normally trained models, attacking defense models, and ablation study are visualized, respectively. Beside adversarial examples, their attention maps with respect to the ground truth are calculated by Grad-cam [4], which will illustrate the "defocusing" effect and stronger transferability of our FIA as compared to the state-of-the-art attacks. Meanwhile, we also provide some additional experimental results suggested by the reviewers to further demonstrate the effectiveness of our method.*

## A. Attack Normally Trained Models

Adversarial examples and corresponding attention maps in attacking normally trained models are shown in Figures 1-4. The examples are randomly picked from the testing set. Given adversarial examples from different attacking methods, attention maps are calculated based on different target models. Obviously, the proposed FIA (and FIA+PIM) significantly defocuses the target models as compared to the other methods, *i.e.*, the attention maps on our adversarial examples cannot focus on the important object.

## B. Attack Defense Models

In the same token, Figures 5-8 show the adversarial examples and attention maps in attacking defense models. In summary of attacking normally trained and defense models, attention maps on adversarial examples from existing attacks fail to focus on the object of interest if the target model is just the source model, while the attention would come back to the object of interest when the target model is different from the source model.

However, attention maps on the adversarial examples from the proposed FIA focus more on non-object regions across different target models. In other words, existing attack methods can only defocus the source model and rarely mislead the other target models, while our method can make the other target models fail to capture the important features of the object in most cases, demonstrating the higher transferability of FIA.

## C. Ablation Study

Recap the loss functions in the ablation study in the main submission,

$$\begin{aligned}
\mathcal{L}_1 &= \sum \left| f_k(x) - f_k(x^{adv}) \right|, \\
\mathcal{L}_2 &= \sum (\Delta_{clean} \odot (f_k(x) - f_k(x^{adv}))), \\
\mathcal{L}_3 &= \sum (\Delta \odot (f_k(x) - f_k(x^{adv}))),
\end{aligned}$$

where $\mathcal{L}_1$ optimizes the feature divergence without constraints like most of the baseline methods, $\mathcal{L}_2$ uses non-aggregate gradient $\Delta_{clean}$ (*i.e.*, gradient from the original clean image), and $\mathcal{L}_3$ is equivalent to our proposed loss using aggregate gradient. Figures 9-10 illustrate the adversarial examples and corresponding attention maps using different losses.

## D. Comparison with other SOTA

Table 1 compares our method to other SOTA (*i.e.*, SI-NI-* [3]). Our FIA performs better than SI-NI-FGSM, SI-NI-TIM, SI-NI-DIM and similar to SI-NI-TI-DIM which combines multiple SOTA techniques that are not adopted in FIA. If adapting our method to SI-NI-TI-DIM (*i.e.*, FIA+SINITIDIM), it achieves the best performance, improving SI-NI-TI-DIM by over 10% on average, agree with the results in the main submission.

## E. Performance on stronger defense models

Table 2 compares different methods against stronger defense models, *i.e.*, the top-3 defense solutions from the NIPS 2017 adversarial competition. We can observe that adapting our method to existing attacks will significantly improve the transferability, *i.e.*, FIA+SINITIDIM, which improves the success rate of SINITIDIM by around 18% on average, still aligning with the conclusion in the main submission.

## F. Time consuming

In the experiments, our method conducts 30 iterations to obtain the aggregate gradient and 10 iterations for optimization. Since the aggregate gradient can be calculated in parallel/batch, running 40 iterations in total could cost similar time as others iterate 10 times. For a fair comparison in terms of computational complexity, let the baselines iterate 30+10=40 times as shown in Table 3, where our method still outperforms the others. Note that baselines tend to degrade with more iteration because of overfitting, while ours benefits from increasing iteration due to the proposed aggregate gradient as shown in the main submission. Thus, our method would significantly extend the performance upper boundary if adapted to others.

## G. Effect of other transformations

We further conduct other transformations, *e.g.*, Gaussian noise, Median filter, Gaussian smooth, and JPEG compression. Their success rates on the target model Inception-V4 (the source model is Inception-V3) are 65.8%, 62.5%, 60.2%, and 59.8%. By contrast, we adopt random mask that achieves 83.5%.

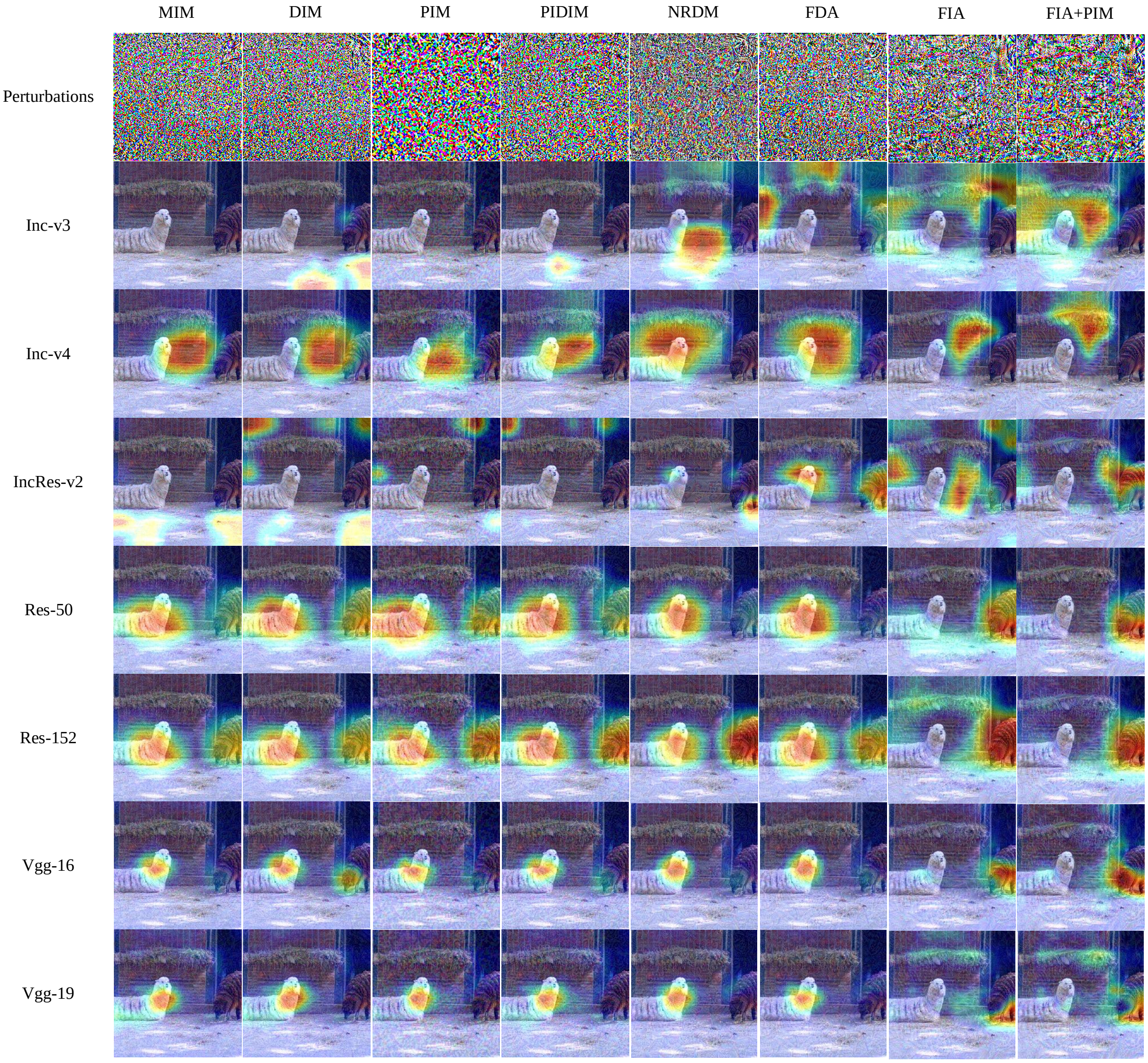

Figure 1. Adversarial examples from different attacks (*i.e.*, the row headers), and the first row shows corresponding perturbations. For the rest rows, the overlaying heat maps are attention from different target models (*i.e.*, the column headers). In this experiment, the source model is Inc-v3.

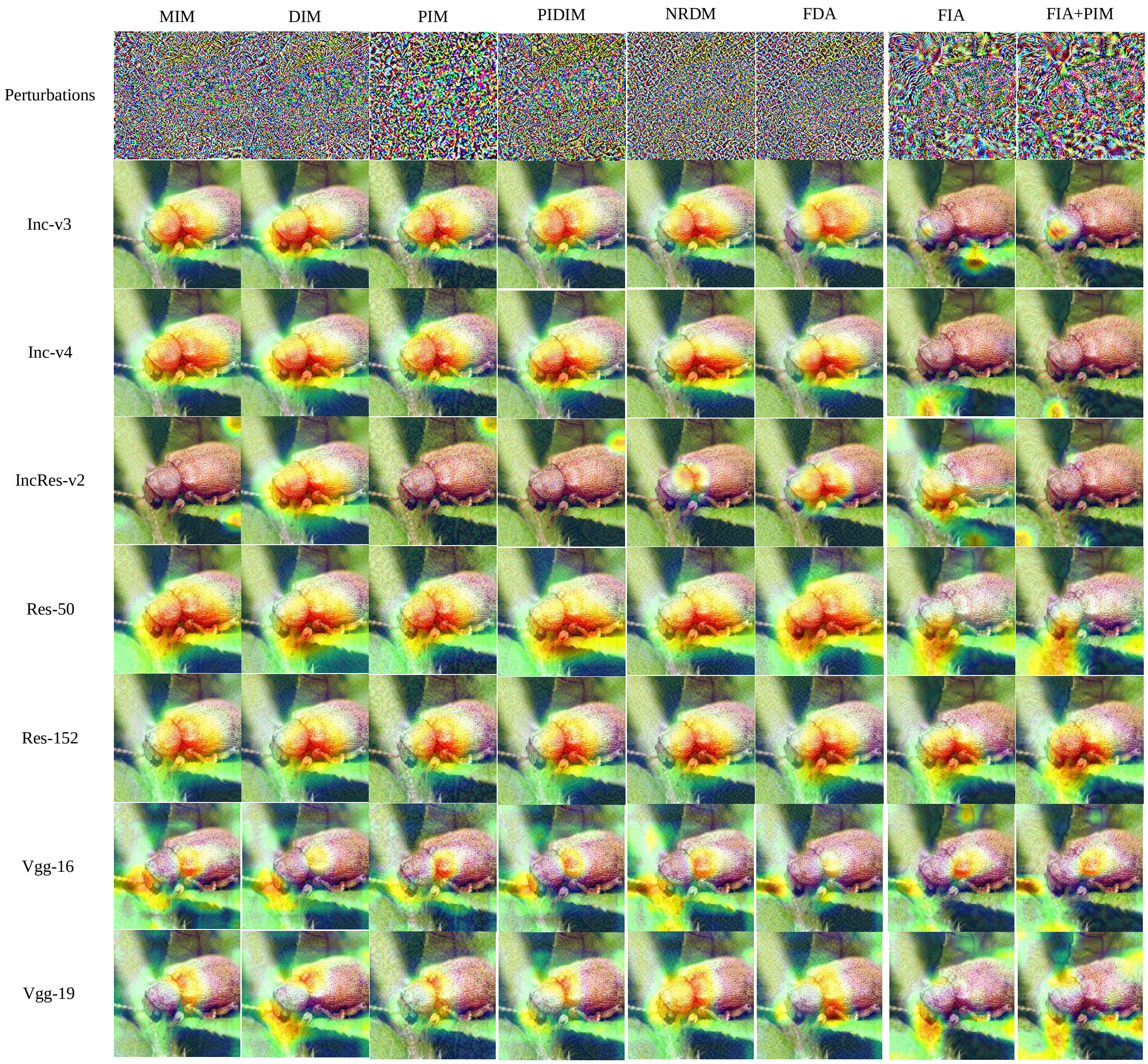

Figure 2. Adversarial examples from different attacks (*i.e*., the row headers), and the first row shows corresponding perturbations. For the rest rows, the overlaying heat maps are attention from different target models (*i.e*., the column headers). In this experiment, the source model is IncRes-v2.

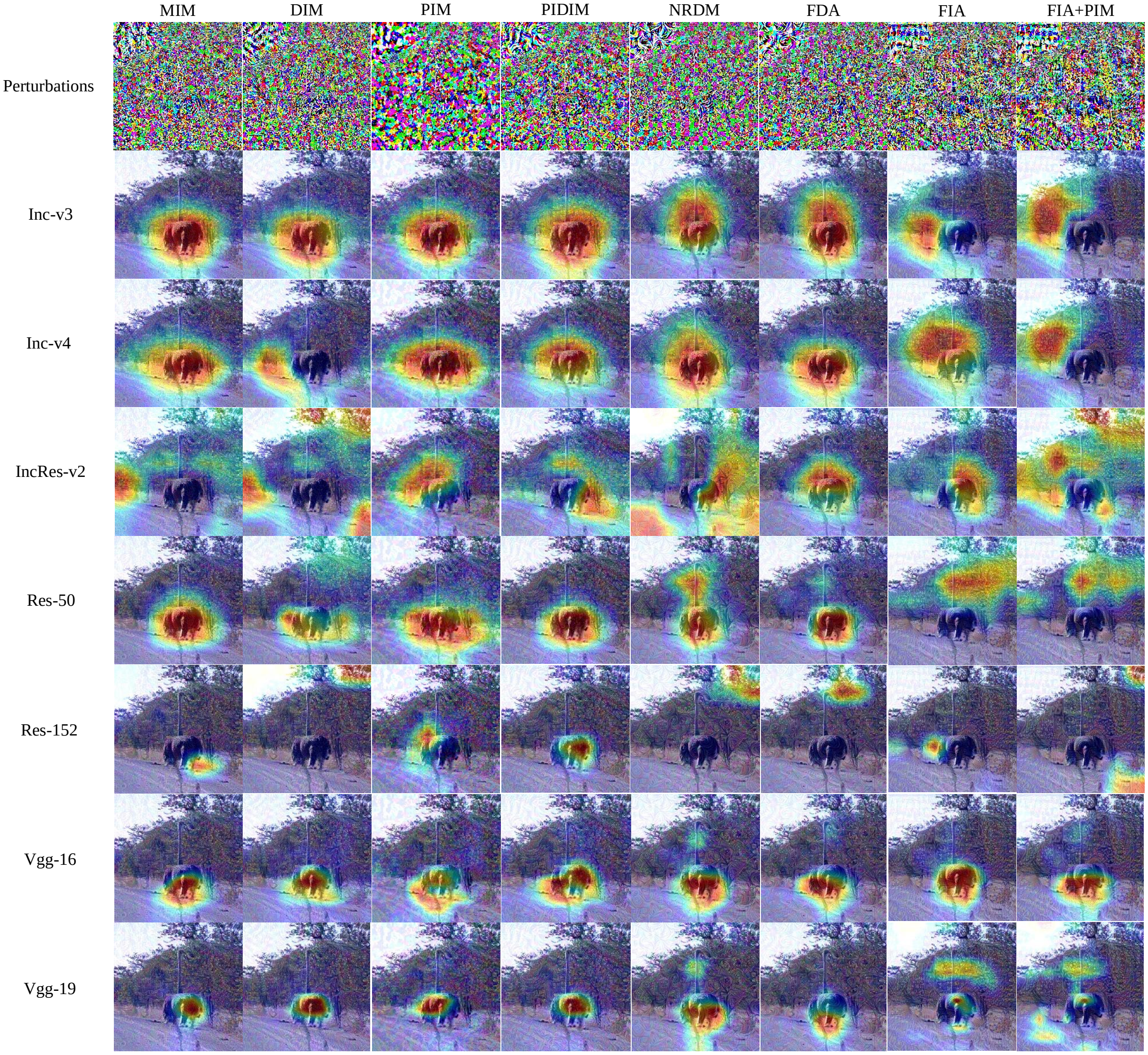

Figure 3. Adversarial examples from different attacks (*i.e.*, the row headers), and the first row shows corresponding perturbations. For the rest rows, the overlaying heat maps are attention from different target models (*i.e.*, the column headers). In this experiment, the source model is Res-152.

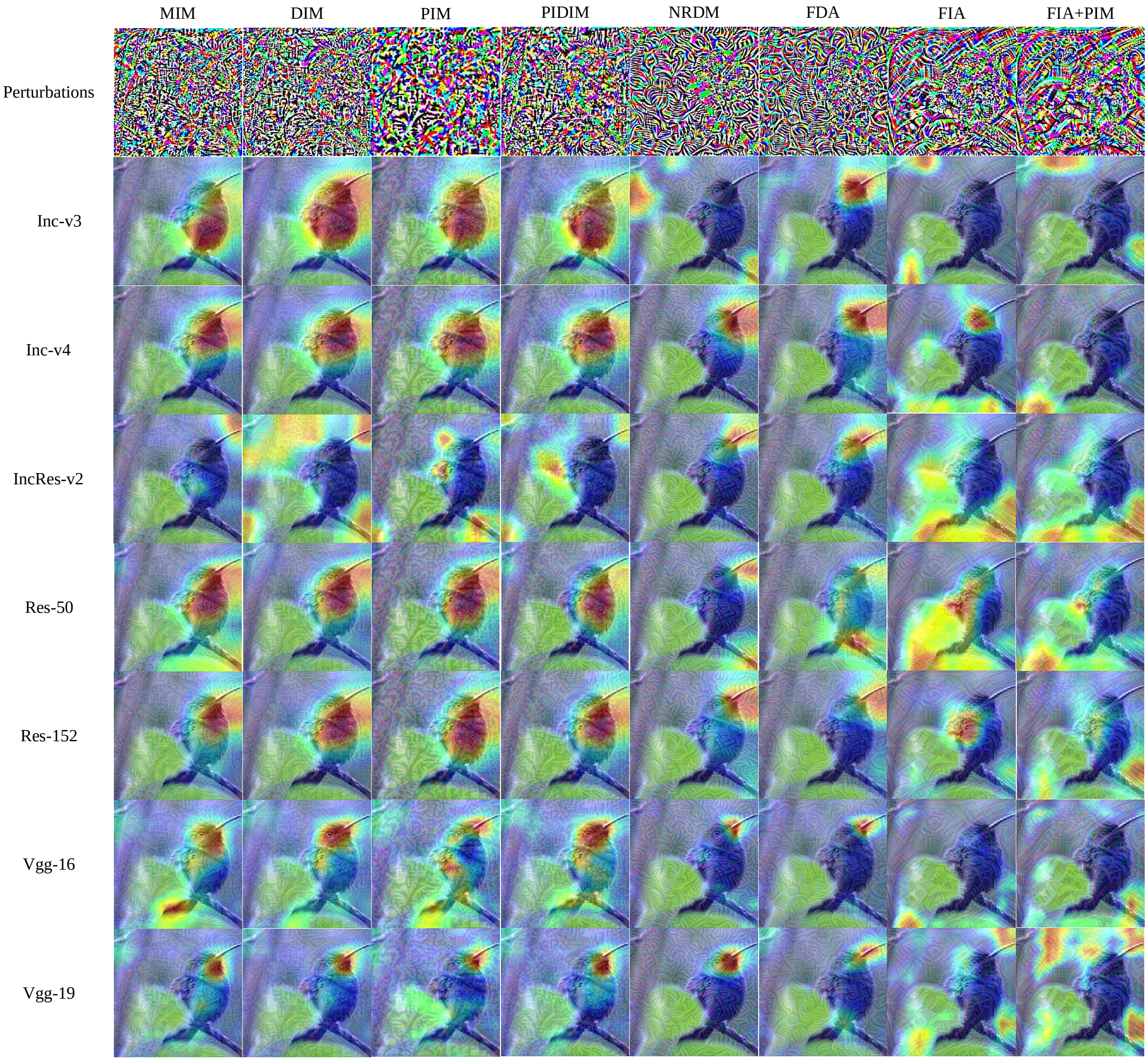

Figure 4. Adversarial examples from different attacks (*i.e.*, the row headers), and the first row shows corresponding perturbations. For the rest rows, the overlaying heat maps are attention from different target models (*i.e.*, the column headers). In this experiment, the source model is Vgg-16.

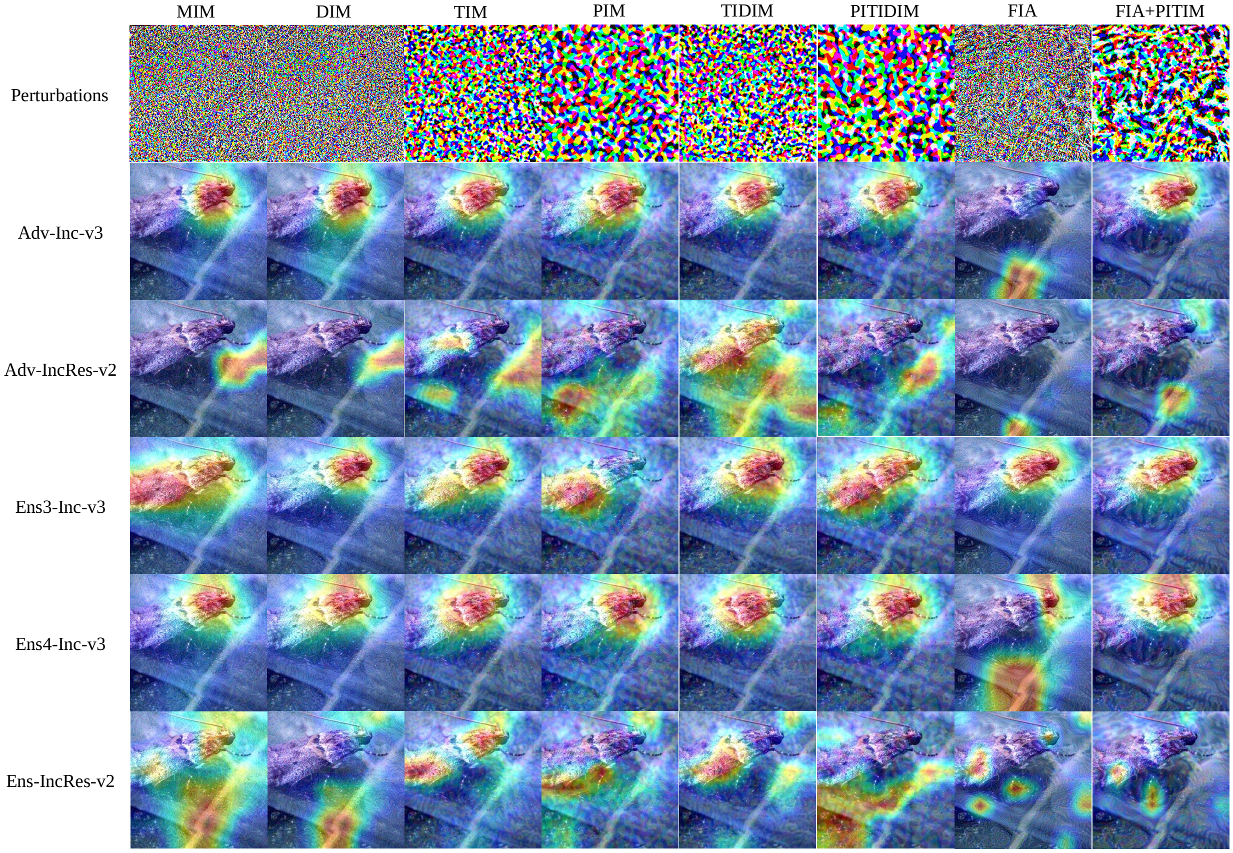

Figure 5. Adversarial examples from different attacks (*i.e*., the row headers), and the first row shows corresponding perturbations. For the rest rows, the overlaying heat maps are attention from different target models (*i.e*., the column headers). In this experiment, the source model is Inc-v3.

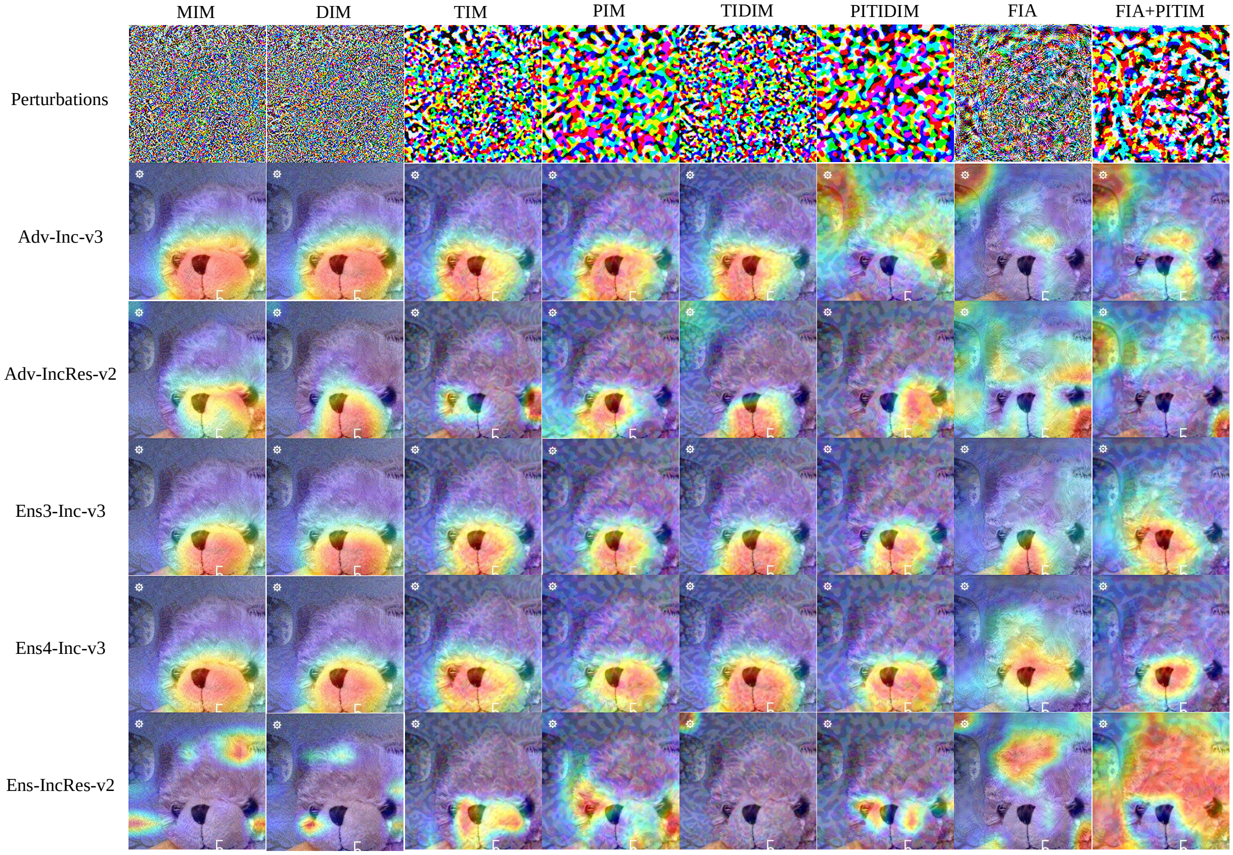

Figure 6. Adversarial examples from different attacks (*i.e*., the row headers), and the first row shows corresponding perturbations. For the rest rows, the overlaying heat maps are attention from different target models (*i.e*., the column headers). In this experiment, the source model is IncRes-v2.

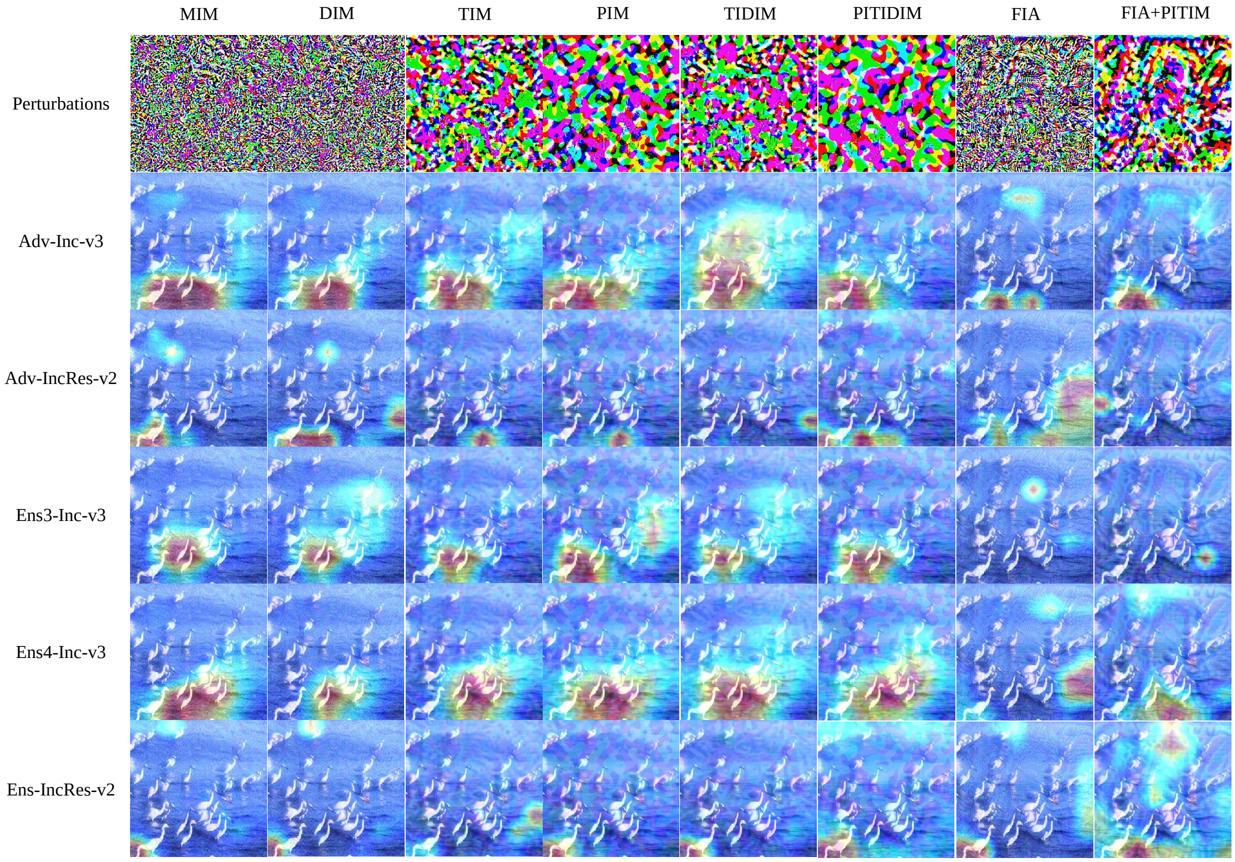

Figure 7. Adversarial examples from different attacks (*i.e*., the row headers), and the first row shows corresponding perturbations. For the rest rows, the overlaying heat maps are attention from different target models (*i.e*., the column headers). In this experiment, the source model is Res-152.

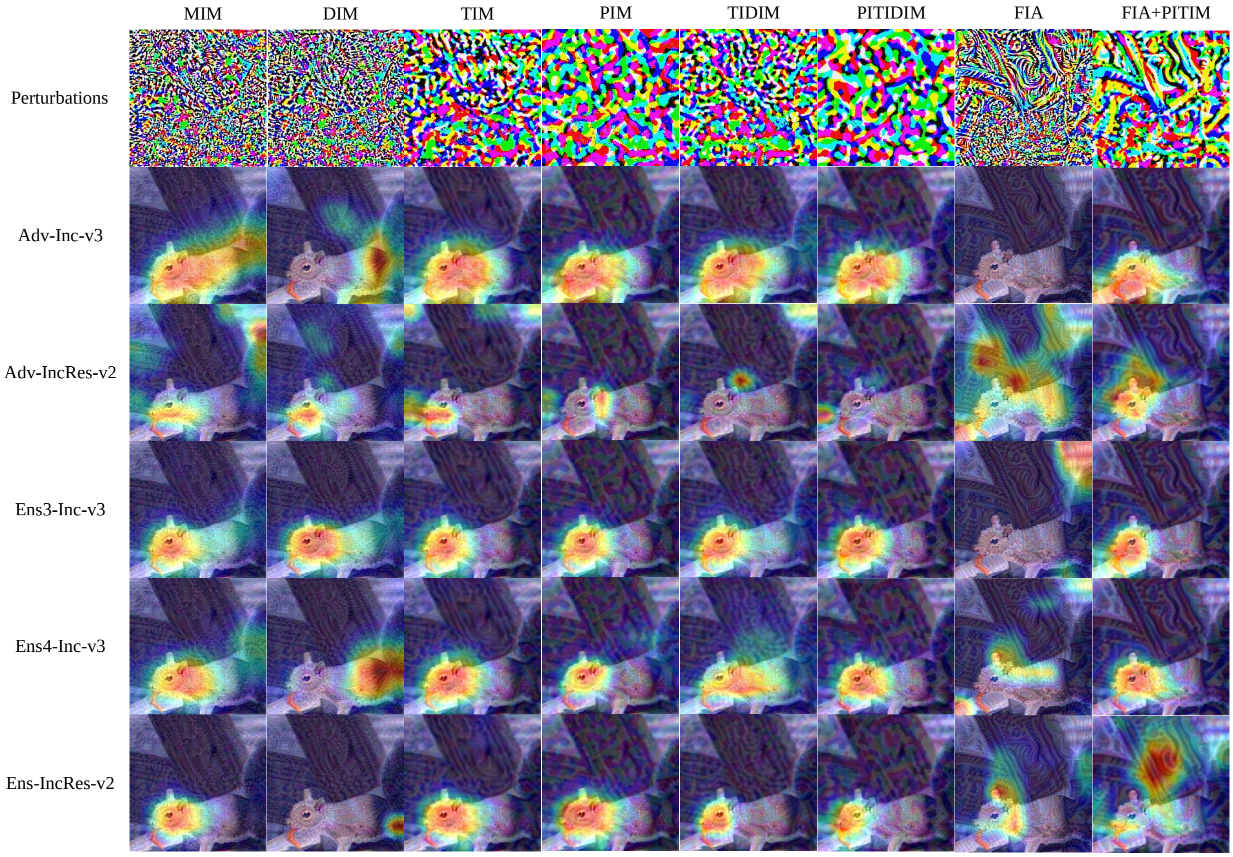

Figure 8. Adversarial examples from different attacks (*i.e.*, the row headers), and the first row shows corresponding perturbations. For the rest rows, the overlaying heat maps are attention from different target models (*i.e.*, the column headers). In this experiment, the source model is Vgg-16.

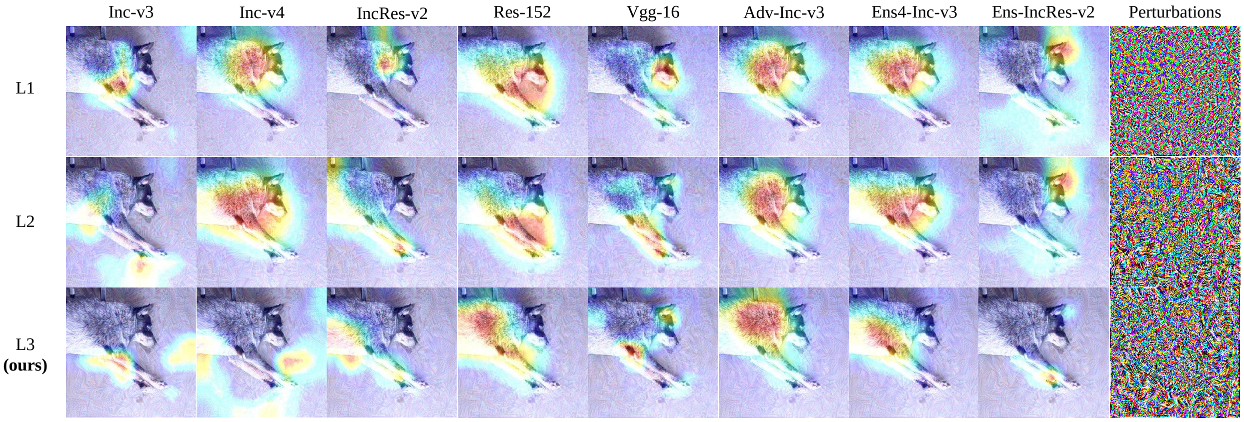

Figure 9. Adversarial examples from different losses (*i.e.*, the column headers), and the last column shows corresponding perturbations. For the rest columns, the overlaying heat maps are attention from different target models (*i.e.*, the row headers). In this experiment, the source model is Inc-v3.

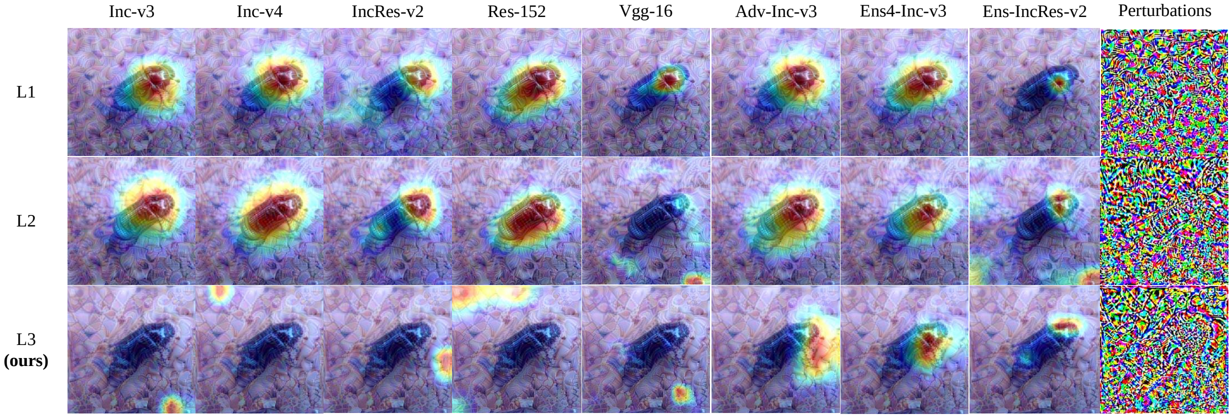

Figure 10. Adversarial examples from different losses (*i.e.*, the column headers), and the last column shows corresponding perturbations. For the rest columns, the overlaying heat maps are attention from different target models (*i.e.*, the row headers). In this experiment, the source model is Vgg-16.

Table 1. Success rate of differnet methods. The source model is Inception-V3, "*" indicates white-box attack.

| Attack | Inc-v3 | Inc-v4 | IncRes-v2 | Res-50 | Res-152 | Vgg-16 | Vgg-19 |
|---|---|---|---|---|---|---|---|
| SI-NI-FGSM | **100.0%*** | 76.7% | 74.0% | 58.1% | 52.5% | 58.7% | 60.9% |
| SI-NI-TIM | **100.0%*** | 75.4% | 71.4% | 65.1% | 59.3% | 70.0% | 69.3% |
| SI-NI-DIM | 98.8%* | 82.2% | 80.0% | 66.7% | 60.3% | 69.4% | 68.5% |
| SI-NI-TI-DIM | 98.5%* | 82.4% | 79.0% | 73.4% | 67.4% | 80.7% | 77.4% |
| FIA | 98.3%* | 83.5% | 80.6% | 70.4% | 64.9% | 71.4% | 73.3% |
| FIA+SINITIDIM | 97.9%* | **89.8%** | **88.1%** | **87.0%** | **85.5%** | **90.8%** | **90.6%** |

Table 2. Success rate of different attacks on stronger defense models. The source model is Inception-V3.

| Defense Models | MIM | DIM | TIM | PIM | TIDIM | PITIDIM | SINITIDIM | FIA | FIA<br>+PITIDIM | FIA<br>+SINITIDIM |
|---|---|---|---|---|---|---|---|---|---|---|
| HGD [2] | 5% | 7.8% | 20% | 23.1% | 31% | 31.3% | 51.0% | 15.4% | 53.3% | **68.1%** |
| R&P [5] | 7.8% | 11.3% | 19.6% | 26.9% | 30.4% | 33.8% | 45.1% | 24.2% | 50.5% | **66.3%** |
| NIPS-r3 [1] | 10.2% | 15.4% | 24.1% | 28.7% | 34.3% | 37.9% | 53.7% | 34.2% | 56.2% | **71.1%** |

Table 3. Success rate of different attacks with the same number of iterations. The source model is Inception-V3.

| Attack | Inc-v3 | Inc-v4 | IncRes-v2 | Res-50 | Res-152 | Vgg-16 | Vgg-19 |
|---|---|---|---|---|---|---|---|
| MIM(40) | **100.0%*** | 37.3% | 35.2% | 33.1% | 27.5% | 39.2% | 36.2% |
| DIM(40) | 96.1%* | 77.3% | 73.6% | 51.3% | 44.5% | 59.0% | 57.6% |
| PIM(40) | **100.0%*** | 51.8% | 47.9% | 49.8% | 44.2% | 61.4% | 59.8% |
| PIDIM(40) | 97.7%* | 78.6% | 74.3% | 53.3% | 47.9% | 59.6% | 60.7% |
| NRDM(40) | 96.9%* | 74.8% | 63.9% | 43.0% | 31.4% | 42.9% | 41.3% |
| FDA(40) | 94.7%* | 57.3% | 46.8% | 33.6% | 26.7% | 36.8% | 34.4% |
| FIA | 98.3%* | **83.5%** | **80.6%** | **70.4%** | **64.9%** | **71.4%** | **73.3%** |

## References

[1] Third place of nips 2017: Defense against adversarial attack. https://github.com/anlthms/nips-2017/tree/master/mmd. 11

[2] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1778–1787. IEEE Computer Society, 2018. 11

[3] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020. 1

[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 1

[5] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 11