ENHANCING THE TRANSFERABILITY OF ADVERSARIAL ATTACKS VIA SCALE ENSEMBLE

Anonymous authors
Paper under double-blind review

ABSTRACT

There is a line of works on adversarial example generation in computer vision, which makes deep learning suffer a lot. Driven by the transferability decrease among models with different input sizes, we present a novel attack method by using a scale input ensemble framework to enhance the transferability of adversarial images, which is named Scale Ensemble Method(SEM). Our method can preserve the characteristic texture of the original image via zooming the surrogate model's input in and out in a specific sequence during generating adversarial examples. The superior texture feature highlights the important attacking region and increases the diversity of adversarial perturbations for assisting a more aggressive attack. The experiments on ImageNet show that our method successfully mitigates the gap of transferability between models with different input sizes and achieves about 8% higher success rate comparing with the state-of-the-art input transformation methods. And we also demonstrate that our method can integrate with existing methods and bypass a variety of defense methods with over 90% success rate.

1 INTRODUCTION

Adversarial examples are widely used in different scenarios include image classification (Goodfellow et al. (2014)),image generation (Chen et al. (2021b)), video classification (Chen et al. (2021a)), *etc* for testing the venerability of deep neural networks. Those adversarial examples can generate closely with the target model (white-box attack) or do not rely on the target model (black-box attack). For the safety of DNN, both adversarial attack and defense have been attracting a lot of attention in both academia and industry.

Despite the attack methods have been proposed differently for different scenarios, another important direction is producing more transferable adversarial examples. While, we found some features when we explore the adversarial transferability. We attack the pretrained models in Pytorch using size 3x224x224 while the Inception(Szegedy et al. (2016)) uses the size 3x299x299, and the transferability between these two kinds of models usually suffers decreasing. This reminds us that almost all the previous works craft adversarial examples in a model-specific image size, which makes the adversarial examples sensitive to the image interpolation method (Quiring et al. (2020)).

Based on the above, we start to rescale adversarial examples in various sizes and ensemble them together to omit the scale feature that the target model learned,, shown in Fig. 1, which we called **Scale Ensemble Method** (SEM). Note that SEM is different from the previous adversarial method SIM(Lin et al. (2020)), which scales the pixel values of images while ours change the image size. The method we used to change the input size is inspired by the excellent work DeepDreams[1], which feeds the networks with images of different sizes to enhance rich features that neural networks have learned. Furthermore, the attention maps of images with various scales also show that different scales of images can produce different attention maps. This indicates that our method can contain more different attention areas which may reflect in the adversarial perturbation generating process. More importantly, compared with other input-transformation methods, we do not import other distribution noises for increasing the input diversity, but only use the different scales of the inputs to capture the various features related to the texture and segmentation of the original images.

[1] `https://www.tensorflow.org/tutorials/generative/deepdream`

We evaluate our approach on four different image classification models pretrained on ImageNet(Deng et al. (2009)), and compare our method with three state-of-the-art input transformation adversarial methods. The results show our proposal outperforms the other methods 5% to 10% success rate across all models. And we also integrate our method with others to propose a new SOTA attack method. Experiments on clean and adversarially-trained models both demonstrate the effectiveness of our new ensemble method. Finally we test our attack on ten adversarial defense methods. The results show that SEM achieves an average of 90.5% success rate on four models against ten defense methods, which is 5.8% higher than the previous best method.

The contributions of this paper are as follows:

- Different from other input transformation-based methods, our work can maintain the input texture to form more robust adversarial attacking features. We reveal that the image scale can be a quite important factor for helping adversarial examples generation.
- Our work is outperforming other attack methods, and can also joint with other attack algorithms to promote attack transferability. Even in defense settings, our method can also maintain efficient attack performance.
- We provide an explanation of the mechanism of our attacking method, which suggests that the diversity of gradient is important for generating more transferable adversarial examples.

2 Related Works

2.1 Adversarial Attacks

There are a lot of methods that have been proposed for adversarial examples generation. Adversarial attacks can be classified as white-box attacks and black-box attacks. In the white-box setting, they can also be classified as gradient optimization based methods and input transformation based methods. In this work, we mainly focus on promoting the attack performance of input transformation based methods. Here, we provide a brief introduction to a number of recently introduced white-box attacks in the literature.

Gradient Optimization Based Attack. Szegedy et al. (2013), Goodfellow et al. (2014) and Kurakin et al. (2018) starts the adversarial examples in deep learning. Dong et al. (2018) proposed MI-FGSM which integrate momentum into the iterative attack and lead to a higher transferability for adversarial examples. Translation Invariant Method (TIM) was also proposed by Dong et al. (2019) to evading the defense models with the Gaussian Smooth to smooth the local gradients during iterations. Lin et al. (2020) claim that Nesterov accelerated gradient is superior than momentum, they adapt Nesterov accelerated gradient as a gradient optimization method to improve the transferability of adversarial examples. Wang & He (2021) proposed a method called variance tuning to enhance the class of iterative gradient based attack methods, they use the gradient variance of the previous iteration to tune the current gradient so as to stabilize the update direction and escape from poor local optima.

Input Transformation Based Attack. Xie et al. (2019) optimize the adversarial perturbations over the diverse transformation of the input image at each iteration. The transformations include the random resizing and the random padding. Diverse Input Method (DIM) can be naturally integrated into other gradient-based attacks to further improve the transferability of adversarial examples. Lin et al. (2020) modify the input image with the scale factor which is divide the value of image pixels by 2^i. The approach changes the input images a lot to avoid "overfitting" on the white-box model. Wu et al. (2018) utilizes the variance reduced gradient to generate more transferable adversarial examples, says an averaged gradient can removes the local fluctuation compared with others.

2.2 Defense

Adversarial defense develops together with attacks. One of the most famous approach to defending against adversarial attacks is Adversarial Training(AT, Goodfellow et al. (2014)), which trains deep neural network(DNN) with adversarial examples to enhance the model's robustness against adversarial attacks. Moreover, Tramèr et al. (2018) proposed Ensemble Adversarial Training(EAT), a technique that augments training data with perturbations transferred from other models. EAT exhibit much stronger robustness due to the higher transferability of its training adversarial examples.

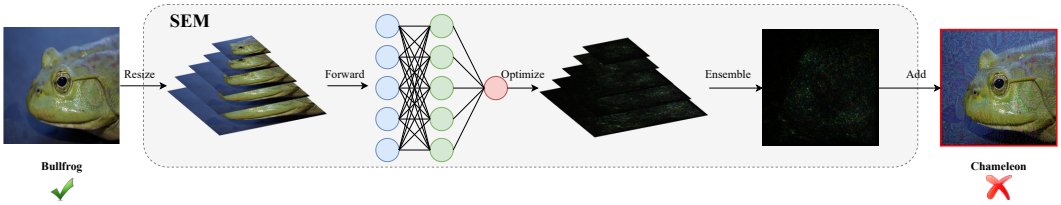


Figure 1: The process to generate adversarial examples via **SEM**. The input image first be transformed into different image sizes. Then, through the target model, we calculate the related gradients of those inputs. Finally, we integrate the related gradients to generate the adversarial perturbation.

Besides adversarial training, there is a major class of adversarial defenses processing the input images to achieve robustness. Guo et al. (2018) use JPEG compression and Total Variance Minimization (TVM) to compress the perturbed images. Similarly, Liu et al. (2019) propose a DNN-oriented JPEG compression method named Feature Distillation (FD). Xie et al. (2018) mitigate the impact of adversarial perturbation via randomly Resize and Padding (R&P). Xu et al. (2018) squeeze the perturbed images via reducing the bit color(Bit-Red) of input images. Cohen et al. (2019) use Random Smoothing (RS) to recover the perturbed accuracy. Prakash et al. (2018) use the pixel deflection (PD) to redisrtibute the input images. Liao et al. (2018) train a denoising model guided by the high-level representation of DNN (HGD). Naseer et al. (2020) similarly train a purifier network through minimize the mid-level feature distance between clean and adversarial images, which is called Neural Representation Purifier (NRD). Besides, Mustafa et al. (2019) propose a novel defense strategy that uses the image super-resolution (SR) to destroy the adversarial perturbation.

In this paper, we choose both AT and EAT to evaluate our attacks. And all of the above defense methods are used to test the effectiveness of our method.

3 METHODOLOGY

In this paper, we propose a novel input transformation method that exhibit the features of images at different scales named **Scale Ensemble Method**(SEM). Typical examples are shown in Fig. 1. We provide the detail description of our approach in the next.

Given an image x with ground-truth label y, and a classification model $f(x) : \mathcal{X} \rightarrow \mathcal{Y}$, we want to generate an adversarial example x^{adv} that fools the classifier and keep the perturbation unperceptual simultaneously. Denote J is the cost function for target classifier f, the goal of crafting an adversarial example x^{adv} can be described as

$$\arg\max_{x^{adv}} J(x^{adv}, y), \quad s.t. \ \|x - x^{adv}\|_\infty \leq \epsilon \tag{1}$$

A group of methods have been proposed to solve the optimization problem in Eq.1 based on FGSM(Goodfellow et al. (2014)). We take MI-FGSM(Dong et al. (2018)) as an example. Compared with the original single-step method FGSM, MI-FGSM is extend to an iterative version and uses momentum item to boost the optimization process. The iterative update rule of MI-FGSM is

$$\begin{aligned} g_{t+1} &= \mu \cdot g_t + \frac{\nabla_x J(x_t^{adv}, y^{true})}{\left\|\nabla_x J(x_t^{adv}, y^{true})\right\|_1}, \\ x_{t+1}^{adv} &= \text{Clip}_x^{\epsilon}\{x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1})\}, \end{aligned} \tag{2}$$

where $\text{sign}(\cdot)$ function restricts the perturbation in the L_∞ norm bound, $\text{Clip}_x^{\epsilon}(\cdot)$ function restricts generated adversarial examples to be within the ϵ-ball of x, $\nabla_x J$ denote the derivative function of J to x, μ is decay rate and g_t is the accumulated gradient at iteration t. Based on MI-FGSM, Lin et al. (2020) uses the Nestrov(Nesterov (1983)) to accelerate the convergency, which is named NI-FGSM.

3.1 INPUT TRANSFORMATION METHODS

Although the development of FGSM makes the optimization faster and preciser, the transferability of adversarial examples they generate is still very low. A series of input transformation methods have been proposed to mitigate the problem. And we will introduce several of them here briefly.

Variance Reduced Method(VRM). VRM propose that the gradients shatter in the local landscapes of DNN's input space, and they suggest to smooth the landscape with Gaussian mollifier, the true gradient to update the adversarial examples at the point x is calculated by

$$G_\sigma(x) = \mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2)}[g(x+\xi)] \tag{3}$$

where σ is variance of sampled Gaussian noise. In the paper, they sample the noise m times and average the gradient as the gradient expectation ta x.

Diverse Input Method(DIM). Xie et al. (2019) think that pure optimizations like MI-FGSM are easy to overfit the specific networks. So they propose to randomly resize the pad the original images with a given probability.

Scale Invariant Method(SIM). Lin et al. (2020) propose to optimize the adversarial perturbation over the scale copies of the input image, where the transformation they use is $S_i(x) = x/2^i$, i.e. simply dividing the pixel values of x with 2^i.

Although these input transformation methods do improve the transferability of adversarial examples, they all generate adversarial examples in a model-specific size. Once the adversarial examples are generated, the only way to attacking a target model with different input size is resizing. However, this usually doesn't work well(see detail in Sec.4.2).

Another problem of existing input transformation methods is *gradient diversity*. The key point of input transformation methods is to avoid overfitting in a single model. To be more specific, they want to find a generalizable gradient that can fit all the potential victim models, which is rarely possible due to the imperfect performance of target models. Existing methods like DIM and VRM augment the input data in a random manner, but the gradients of these methods rarely change with the transformations. These approaches have small gradient diversity so that once the white-box model gives wrong attention the update gradient is totally wrong. SIM is a little different that augments inputs via dividing pixel value. The gradients of SIM is much diverse but not concentrated when the scale factor i is very large. In other words, the model often gives the wrong gradients when i is large, and the accumulated gradients of SIM have more bias. The detail of above analysis of gradient can be seen in Sec.4.6.

Based on above discussion, we propose our Scale Ensemble Method that augments the inputs in various image scales and keep gradients more diverse but concentrate on the accurate areas.

3.2 Scale Ensemble Method

Denote T_α is the transformation which resizes the image with size s into αs. The interpolation methods we use in this paper is the Nearest Neighbour Interpolate for the its low computation cost.

We introduce *scale range* l and r where we want our image sizes to range and the *scale number* N. We first generate N evenly spaced resize factors over scale interval $[l, r]$. The new scale α_i of input image can be described as

$$\alpha_i = \text{Ceil}\{[l + i(r-l)] \cdot s\}, \quad i = 0, 1, ..., N-1. \tag{4}$$

A group of transformation functions $\{T_{\alpha_i}\}_{i=0}^{N-1}$ could be created via Eq.4.

During each attack iteration, the adversarial examples x_t^{adv} will be transformed into N different scales $\{T_{\alpha_i}(x_t^{adv})\}_{i=0}^{N-1}$ and then fed into the classification model f to compute the gradients. The update gradient at iteration t can be computed by the following formulation:

$$\hat{g}_t = \frac{1}{N}\sum_{i=1}^{N} \nabla_x J(T_{\alpha_i}(x_t^{adv}), y^{true}), \tag{5}$$

combining with MI-FGSM to form SE-MI-FGSM, the gradient update rule is

$$\begin{aligned} g_{t+1} &= \mu \cdot g_t + \frac{\hat{g}_t}{\|\hat{g}_t\|_1}, \\ x_{t+1}^{adv} &= \text{Clip}_x^{\epsilon}\{x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1})\}. \end{aligned} \tag{6}$$

Through updating the adversarial examples T times, we can get the final adversarial example x_T^{adv}. The update gradient of SEM is much more diverse compared with other input transformation method

because the each layer of target network perceives different area of the original images, which plays the role like model ensemble, and on the other hand we do not import any other distribution noise so the original texture information can be maintained at a high level.

Note that here our gradient accumulation methods can be combined with other optimization-based attacks like NI-FGSM too. Besides, SEM can be combined with not only gradient constrained attacks but also any other input transformation methods.

4 EXPERIMENTS

In order to validate the effectiveness of our proposed method, we conduct some experiments on ImageNet (Russakovsky et al. (2015)). The construction of this section is as follows: we first introduce the experiment setup including models, dataset, baselines, and the hyper-parameters. Next, we show the comparison of our SEM with other input transformation methods through untarget attacks with a single model. Then we combine a series of previous attacks to introduce the state-of-the-art transfer attack, which outperforms the previous single method attacks with a large margin, and also in this part, we confirm that our SEM can integrate the SOTA and enhancing the transferability further. After that, we test our methods with 10 adversarial defense methods and demonstrate the robustness of the proposed attack. Next we conduct some ablation studies to further validate the effectiveness. Finally we give some analysis of why our method works better than others.

4.1 SETUP

Models. We select 4 common-used ImageNet pre-trained models Inception-v3 (Inc-v3), Inception-Res-v2(IncRes-v2)(Szegedy et al. (2016)), ResNet50(Res-50)(He et al. (2016)) and DenseNet121 (Den-121) (Huang et al. (2017)) as our white-box models to craft adversarial examples. Note the basic input size is 299 for Inception-v3, Inception-ResNet-v2, and 224 for the other models in PyTorch. As for evaluation, we add another two adversarially trained models AdvInception-v3 (Inc-v3$_{adv}$) and Inception-Res-v2$_{ens}$ (IncRes-v2$_{ens}$) (Tramèr et al. (2018)) into our evaluation model list besides the above four models.

Dataset. We follow the setting of papers (Wang & He (2021)) and randomly pick 1000 clean images from the ILSVRC 2012 validation set (Russakovsky et al. (2015)), which are separated from 1000 classes of ImageNet and mostly classified correctly by the evaluation models.

Baselines. In this paper, we implement untarget attacks with two common-used optimization methods MI-FGSM (Dong et al. (2018)) and NI-FGSM (Lin et al. (2020)). We compare SEM with three other input transformation methods DIM, SIM and VRM previously mentioned. The name of the integrated method starts with the input transformation method followed by the basic optimization method like SE-MI-FGSM, DI-NI-FGSM, etc.

Hyper-parameters. We follow the setting of (Wang & He (2021)) with maximum perturbation as $\epsilon = 16$, number of iterations $T = 10$, and step size α is set as $\epsilon/T = 1.6$. The decay rate of momentum-based optimization methods MI-FGSM and NI-FGSM is 1.0. The diverse input probability of DIM is setting as 0.5, kernel size of TIM is 7×7, and scale of copies of SIM is 5(i.e. $i = 0, 1, 2, 3, 4$). For VT, we set $N = 20, \beta = 1.5$. The hyper-parameters m, σ for VR is 20 and 20 following the original paper. Finally for SEM, we set the scale number $N = 50$, and l and r are 0.7 and 1.3.

4.2 SINGLE MODEL ATTACKS

We first conduct the experiment of attacking a single model with four different input transformation methods. In detail, we conduct the untarget adversarial attacks with 4 models previously metioned. All the success rates are calculated with 1000 adversarial examples in this paper. The results are shown in Tab.1, where the row indicates the attack models and the column indicates evaluation models.

According to Tab.1 we can first conclude that almost all white-box attacks succeed except for DenseNet, where our method achieves the success rate of 97.4% compared with 93.3% that the

Table 1: The success rate (%) of single model attacks. The rows indicate the attack models while the columns are evluation models.

Model	Attack	Inc-v3	IncRes-v2	Res-50	Den-121	Inc-v3$_{adv}$	IncRes-v2$_{ens}$
Inc-v3	VR-MI-FGSM	99.9*	80.1	68.9	68.9	**55.5**	**36.0**
	DI-MI-FGSM	99.5*	64.0	49.6	49.6	30.3	12.0
	SI-MI-FGSM	**100.0***	68.9	56.6	56.6	38.5	17.0
	SE-MI-FGSM	**100.0***	**87.2**	**74.8**	**74.8**	51.4	31.5
	VR-NI-FGSM	**100.0***	80.2	66.1	66.1	49.3	29.0
	DI-NI-FGSM	99.9*	64.7	49.0	49.0	27.3	9.2
	SI-NI-FGSM	**100.0***	75.8	63.6	63.6	39.9	18.4
	SE-NI-FGSM	**100.0***	**89.8**	**79.0**	**79.0**	**51.3**	**30.0**
IncRes-v2	VR-MI-FGSM	82.1	95.3*	69.0	69.0	58.8	**52.4**
	DI-MI-FGSM	69.2	90.9*	52.0	52.0	31.4	19.2
	SI-MI-FGSM	84.2	**98.6***	67.7	67.7	53.5	43.9
	SE-MI-FGSM	**87.0**	97.5*	**76.3**	**76.3**	**60.1**	49.3
	VR-NI-FGSM	83.2	98.8*	65.3	65.3	49.8	37.3
	DI-NI-FGSM	65.6	97.1*	46.6	46.6	25.3	11.8
	SI-NI-FGSM	85.8	**99.3***	69.9	69.9	52.4	37.2
	SE-NI-FGSM	**88.1**	98.9*	**74.6**	**74.6**	**53.9**	**37.7**
Res-50	VR-MI-FGSM	92.7	91.7	99.8*	99.8	87.7	81.1
	DI-MI-FGSM	84.1	80.5	99.7*	99.7	68.3	53.7
	SI-MI-FGSM	86.0	81.7	**100.0***	**100.0**	71.2	58.5
	SE-MI-FGSM	**98.4**	**97.6**	**100.0***	**100.0**	**92.6**	**86.8**
	VR-NI-FGSM	90.1	89.6	99.9*	99.9	82.0	71.6
	DI-NI-FGSM	79.1	71.8	99.9*	99.9	56.7	37.6
	SI-NI-FGSM	86.3	81.4	**100.0***	**100.0**	68.9	54.1
	SE-NI-FGSM	**98.5**	**97.6**	**100.0***	**100.0**	**91.4**	**83.2**
Den-121	VR-MI-FGSM	89.0	85.7	93.3	93.3*	81.9	72.5
	DI-MI-FGSM	76.6	68.4	81.7	81.7*	57.4	41.4
	SI-MI-FGSM	80.4	72.4	85.7	85.7*	65.7	50.6
	SE-MI-FGSM	**95.6**	**94.1**	**97.1**	**97.1***	**89.2**	**79.4**
	VR-NI-FGSM	85.4	80.8	92.0	92.0*	77.1	66.1
	DI-NI-FGSM	71.5	61.9	79.3	79.3*	50.3	31.6
	SI-NI-FGSM	79.3	70.9	85.8	85.8*	63.5	45.8
	SE-NI-FGSM	**94.6**	**93.4**	**97.4**	**97.4***	**87.5**	**73.8**

others' best. It's notable that success rates of adversarial examples generated by ResNet-50 achieve nearly 100% when transferred to the other three normally trained models.

Next, through comparing the transfer attacks between models with input sizes 299×299 (Inc-v3 and IncRes-v2) and 224×224 (Res-50 and Den-121), we can find that the transferability downgrades obviously when compared with the same input sizes. On the other hand, comparing SEM with others when the input size varies, we can conclude that our method alleviates the transferability decrease phenomenon a lot.

At last, when evaluating the transferability among models of the same size, our method also outperforms the other three models. And in general, we can conclude that SEM's transferability is about 8% higher than VR, which is the previous best input transformation method according to the table.

4.3 Integrated with Other Methods

In this part, we conduct the experiments that prove our SEM can integrate with SOTA transfer-based adversarial attacks and enhance the transferability further. To be more exact, we combine five transferability enhancement methods (VT, TIM, DIM, VR, and SIM) with two basic optimization methods MI-FGSM and NI-FGSM to compose PB-M(N)I-FGSM, which means the previous best methods. After that, we integrate SEM with them to form SE-PB-M(N)I-FGSM. The experiments are also implemented on four attack models and six evaluation models with 1000 images.

As Tab.2 shows, compared with the single-method attacks the combination of previous transfer-based attacks helps enhance the transferability with a large margin. And by integrating our method with the previous best adversarial attack, our method achieves the result that nearly any pair of nor-

Table 2: The success rate (%) of single model attacks integrated with gradient enhancement methods.

Model	Attack	Inc-v3	IncRes-v2	Res-50	Den-121	Inc-v3$_{adv}$	IncRes-v2$_{ens}$
Inc-v3	PB-MI-FGSM	99.8*	89.3	85.7	89.5	87.7	76.1
	SE-PB-MI-FGSM	**100.0***	**95.9**	**93.4**	**95.6**	**92.5**	**84.8**
	PB-NI-FGSM	**100.0***	86.3	82.1	88.7	83.8	70.6
	SE-PB-NI-FGSM	**100.0***	**94.6**	**91.8**	**93.9**	**90.4**	**79.0**
IncRes-v2	PB-MI-FGSM	90.5	97.1*	85.0	89.3	87.4	87.5
	SE-PB-MI-FGSM	**93.1**	**97.8***	**89.0**	**92.6**	**91.5**	**89.7**
	PB-NI-FGSM	**93.2**	99.0*	85.9	90.7	86.8	84.9
	SE-PB-NI-FGSM	92.7	**99.3***	**86.7**	**91.8**	**90.8**	**86.5**
Res-50	PB-MI-FGSM	94.4	90.7	99.7*	97.4	92.0	89.1
	SE-PB-MI-FGSM	**98.6**	**97.9**	**100.0***	**99.6**	**96.6**	**95.4**
	PB-NI-FGSM	93.1	90.2	99.9*	98.3	90.7	86.9
	SE-PB-NI-FGSM	**98.9**	**97.9**	**100.0***	**99.6**	**97.0**	**95.5**
Den-121	PB-MI-FGSM	89.5	84.4	91.6	**100.0***	89.1	85.5
	SE-PB-MI-FGSM	**95.7**	**94.3**	**97.4**	99.9*	**94.4**	**92.3**
	PB-NI-FGSM	89.6	84.3	91.8	**100.0***	88.2	82.8
	SE-PB-NI-FGSM	**96.5**	**94.6**	**97.5**	99.9*	**94.5**	**90.8**

mal models can transfer to each other with an average of about 95% success rate, which outperforms the PB about 5% across all models. When against the adversarially trained models, the proposed SE-PB-M(N)I-FGSMs also have superior performance, whose transfer rate is about 10% higher than the previous best method and achieves 90% on average. Note that all our attacks are implemented with only single models instead of the model ensemble(Wang & He (2021); Dong et al. (2019)), and in general our proposed method can be integrated with other methods and make the success rate of transfer attacks achieves over 90%.

4.4 Attacking Defense Models

Following Wang & He (2021); Dong et al. (2019), we also evaluate the transferability of adversarial exmaples against Inc-Res-v2$_{ens}$ with 10 advanced defense methods to demonstrate the robustness of our attacks. In detail, the defense methods include the top-2 in NIPS-2017 competition HGD (Liao et al. (2018)) and R&P (Xie et al. (2018)), and 8 extra recently proposed defense methods JPEG Compression (Guo et al. (2018)), Feature Distillation (FD) (Liu et al. (2019)), Bit-Reduction (Bit-Red) (Xu et al. (2018)), Total Variance Minimization (TVM) (Guo et al. (2018)), Random Smoothing (RS) (Cohen et al. (2019)), Super Resolution (SR) (Mustafa et al. (2019)), Pixel Deflection (PD) (Prakash et al. (2018)) and Neural Representation Purifier (NPR) (Naseer et al. (2020)).

We test all these defense methods with previously mentioned PB-M(N)I-FGSM and SE-PB-M(N)I-FGSM for their best transferability. And the attacking models keep the same as previous while the evaluating model is defended by various defense methods with Inc-Res-v2$_{ens}$ as backbone following (Naseer et al. (2020); Dong et al. (2019)). The experiment results are shown in Tab.3.

From the table, we can observe that for all models and defenses our SE-PB-M(N)I-FGSM outperforms the previous best methods. Comparing with Tab.2 we can find that some of the defense strategies increase the attack success rate instead of decreasing it, and most even don't work on defending both attacks, which takes a great challenge to the failure of defending against adversarial attacks. Note that NRP is the newly proposed defense that trains a generator the purify the adversarial examples using a self-supervised approach, which performs best among all these defenses. However, the proposed method also bypasses it with an average success rate of 79.3% compared with the PB method 71.5%. Finally, as a whole, our proposed method achieves an average success rate of 90.5% which is 5.8% higher than the previous best.

4.5 Ablation Studies

In this section, we conduct the experiment to test the impact of hyper-parameters l, r, and N in our experiments.

Table 3: The success rate (%) of attacks against ten defense methods. The backbone of defense models is IncRes-v2$_{ens}$.

Model	Attack	JPEG	FD	Bit-Red	TVM	HGD	R&P	RS	SR	PD	NRP
Inc-v3	PB-MI-FGSM	79.8	77.9	87.6	85.7	76.0	78.5	76.2	85.1	83.9	74.1
	SE-PB-MI-FGSM	**88.1**	**87.0**	**91.1**	**91.2**	**84.7**	**87.8**	**83.3**	**90.3**	**89.9**	**77.8**
	PB-NI-FGSM	76.3	72.1	82.0	81.1	70.7	71.8	71.1	79.3	78.4	68.0
	SE-PB-NI-FGSM	**84.8**	**81.4**	**88.9**	**87.7**	**79.0**	**83.5**	**76.9**	**86.8**	**86.1**	**75.0**
IncRes-v2	PB-MI-FGSM	89.5	89.0	89.5	89.4	89.5	88.4	87.3	88.8	89.0	80.7
	SE-PB-MI-FGSM	**93.8**	**93.3**	**93.4**	**94.3**	**93.7**	**93.5**	**91.9**	**93.4**	**92.8**	**85.1**
	PB-NI-FGSM	88.9	87.4	88.0	88.4	88.4	87.8	85.3	88.7	86.9	75.7
	SE-PB-NI-FGSM	**92.6**	**92.2**	**93.0**	**92.4**	**93.0**	**92.8**	**90.2**	**92.0**	**91.7**	**80.8**
Res-50	PB-MI-FGSM	90.6	89.0	90.9	92.8	90.4	88.8	89.6	90.4	90.6	81.7
	SE-PB-MI-FGSM	**95.6**	**95.8**	**95.4**	**96.4**	**95.9**	**95.6**	**92.9**	**94.9**	**94.5**	**87.0**
	PB-NI-FGSM	88.4	88.1	90.4	91.9	89.0	87.0	88.3	88.9	88.9	75.5
	SE-PB-NI-FGSM	**95.3**	**95.7**	**95.7**	**96.0**	**95.9**	**96.2**	**92.8**	**95.0**	**94.5**	**84.9**
Den-121	PB-MI-FGSM	86.4	86.2	88.2	90.0	86.7	84.9	86.5	86.9	87.2	77.1
	SE-PB-MI-FGSM	**92.1**	**91.8**	**92.3**	**94.0**	**93.0**	**92.1**	**91.0**	**92.0**	**91.7**	**81.7**
	PB-NI-FGSM	85.2	85.1	85.9	87.6	85.2	83.6	83.5	84.8	85.3	73.4
	SE-PB-NI-FGSM	**92.3**	**90.7**	**92.1**	**93.8**	**92.5**	**92.0**	**89.6**	**91.9**	**91.5**	**79.7**

Scale range. We first study this parameter by fixing the scale number N=50 and then perform the grid search for these two related parameters l and r. For the best effect of scale ensemble, we assume 1 must be in the spatial range (l, r), which means that the image size must range around the target models' original input size. So we set l searched between 0.3 and 1.0 with 0.1 as the interval, and for r we set 1.0 and 1.8 with the same search interval. The grid search is conducted with SP-MI-FGSM for four different models separately. The results are shown in the first row of Fig. 2, where we plot the 3D surfaces of the average success rate on six evaluation models with l and r change. From the figure, we can observe that nearly all the models acquire the best transferability when $l = 0.7$ and $r = 1.3$. And according to the accurate calculation, we can conclude that the peak point for all models is (0.7, 1.3) except VGG-16, whose peak is (0.6, 1.3). The result also demonstrates the efficiency of our parameters used before.

Scale numbers.

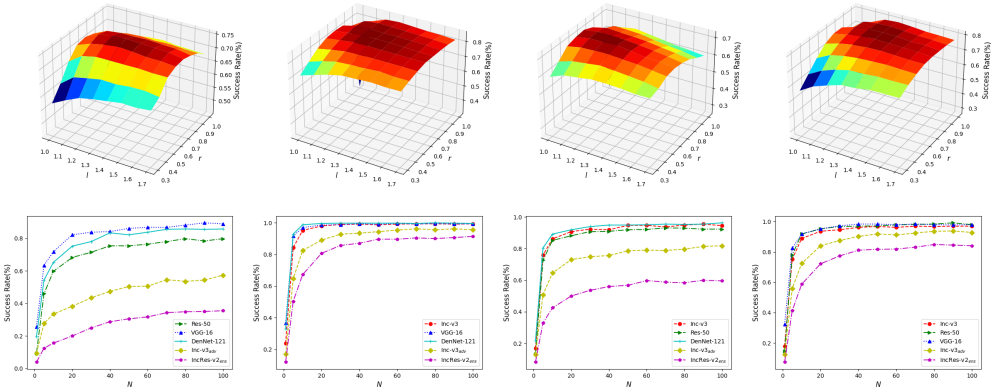
By fixing l and r as previously searched, we change the scale number N between 1 and 100 and then generate 1000 adversarial examples with different Scale numbers separately. All these experiments are conducted via SP-MI-FGSM on four different models. We calculate the success rate of these adversarial examples on all evaluation models except attacking models themselves. Results for four models are shown in Fig. 2.

From the figure, we can see the transferability of Inc-v3 is increasing slowly when $N = 80$ but other models' reach convergency at the point of $N = 40$ clearly. And we can also find that the adversarially trained models converge slower than the normally trained. So, as a conclusion, $N = 50$, which is used in our previous experiments, is the best value for balancing the computational cost and the attacking effects for all models.

4.6 Why is Scale-Ensemble Effective

The purpose of this section is to explain the strength of our method compared with other attacking methods in input diversity. We empirically assume that, by getting the balance of input diversity and attacking regions, our method can generate more robust and unify perturbations for stronger attacks. As illustrated in Fig. 3, we plot five possible attention maps during each iteration with scale changing. The illustration skill comes from Grad-CAM(Selvaraju et al. (2017)).

We can separately see that the first two rows which represents DIM and VRM have similar attention areas while the background images are acquired by random augmentation. This phenomenon shows these two input diverse methods won't change the attention maps a lot, which in other words is unable to diverse the input at a high level. The bottom two raws of Fig.3 show the attention maps with scale increasing. The results show that SIM obviously imports large image-unrelated features so that the attention maps fails concentrating on the importance regions of the attack image. Finally



(a) Inc-v3 (b) Res-50 (c) VGG-16 (d) Den-121

Figure 2: Results of ablation studies on scale range l and r (first row), and on scale number N (sencond row).

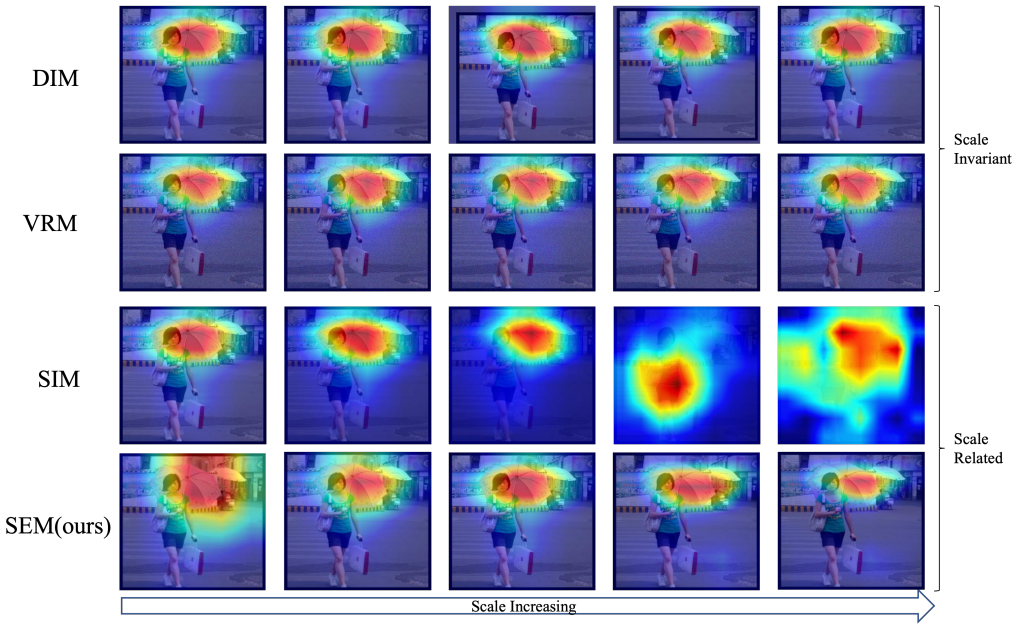


Figure 3: Attention maps change with scale increasing inner a single iteration. Each row indicates a different method. The attention maps of the top two methods which we called *scale-invariant* are not related with the scale changing, while the last two are *scale related*, whose attention maps are related with pixel scales and image scales separately.

as a contrast, our SEM overcomes these two feedbacks and introduces more diverse but correct input images. More illustration results can be seen in A.3.

5 Conclusions

This paper presents a novel method to generating adversarial examples for increasing attack transferability. Furthermore, it can also bypass different defense methods. The core idea is to ensemble the inputs with the different scales that can maintain the most natural texture of original pictures without other noises imported. Experiments compared with three other input transformation methods on four classification models, and ten defense methods validate the effectiveness of the proposed methodology. Our research delivers a key message to the community that the scale of inputs can be an important factor for generating more effective adversarial examples.

REFERENCES

Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Qi Tian. Appending adversarial frames for universal video attack. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3199–3208, 2021a.

Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Bo Zhang. Magdr: Mask-guided detection and reconstruction for defending deepfakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9014–9023, 2021b.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyJ7ClWCb.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *Artificial Intelligence Safety and Security*, pp. 99–112, Jul 2018. doi: 10.1201/9781351251389-8. URL http://dx.doi.org/10.1201/9781351251389-8.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJlHwkBYDH.

Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 860–868. IEEE, 2019.

Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29: 1711–1724, 2019.

Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 262–271, 2020.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady an ussr*, volume 269, pp. 543–547, 1983.

Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8571–8580, 2018.

Erwin Quiring, David Klein, Daniel Arp, Martin Johns, and Konrad Rieck. Adversarial preprocessing: Understanding and preventing image-scaling attacks in machine learning. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pp. 1363–1380, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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 the IEEE international conference on computer vision*, pp. 618–626, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkZvSe-RZ.

Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1924–1933, 2021.

Lei Wu, Zhanxing Zhu, Cheng Tai, et al. Understanding and enhancing the transferability of adversarial examples. *arXiv preprint arXiv:1802.09707*, 2018.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *Proceedings 2018 Network and Distributed System Security Symposium*, 2018. doi: 10.14722/ndss.2018.23198. URL http://dx.doi.org/10.14722/ndss.2018.23198.

A APPENDIX

A.1 ENSEMBLE ATTACKS

We conduct the experiments of model ensemble attacks to demonstrate the effectiveness of SEM further. The ensemble models are Inception-v3, InceptionResNet-v2, ResNet-50 and DenseNet-121 as described before. Results are shown in Tab.4

Table 4: The success rate (%) of model ensemble attacks.

Model	Attack	Inc-v3	IncRes-v2	Res-50	Den-121	Inc-v3$_{adv}$	IncRes-v2$_{ens}$
Inc-v3	PB-MI-FGSM	99.8*	89.3	85.7	89.5	87.7	76.1
	SE-PB-MI-FGSM	**100.0***	**95.9**	**93.4**	**95.6**	**92.5**	**84.8**
	PB-NI-FGSM	**100.0***	86.3	82.1	88.7	83.8	70.6
	SE-PB-NI-FGSM	**100.0***	**94.6**	**91.8**	**93.9**	**90.4**	**79.0**
IncRes-v2	PB-MI-FGSM	90.5	97.1*	85.0	89.3	87.4	87.5
	SE-PB-MI-FGSM	**93.1**	**97.8***	**89.0**	**92.6**	**91.5**	**89.7**
	PB-NI-FGSM	**93.2**	99.0*	85.9	90.7	86.8	84.9
	SE-PB-NI-FGSM	92.7	**99.3***	**86.7**	**91.8**	**90.8**	**86.5**
Res-50	PB-MI-FGSM	94.4	90.7	99.7*	97.4	92.0	89.1
	SE-PB-MI-FGSM	**98.6**	**97.9**	**100.0***	**99.6**	**96.6**	**95.4**
	PB-NI-FGSM	93.1	90.2	99.9*	98.3	90.7	86.9
	SE-PB-NI-FGSM	**98.9**	**97.9**	**100.0***	**99.6**	**97.0**	**95.5**
Den-121	PB-MI-FGSM	89.5	84.4	91.6	**100.0***	89.1	85.5
	SE-PB-MI-FGSM	**95.7**	**94.3**	**97.4**	99.9*	**94.4**	**92.3**
	PB-NI-FGSM	89.6	84.3	91.8	**100.0***	88.2	82.8
	SE-PB-NI-FGSM	**96.5**	**94.6**	**97.5**	99.9*	**94.5**	**90.8**
Ensemble	PB-MI-FGSM	99.3	95.8	97.8	99.5	94.8	90.8
	SE-PB-MI-FGSM	**99.7**	**99.0**	**99.3**	**99.8**	**97.9**	**95.8**
	PB-NI-FGSM	**99.9**	93.1	99.2	99.9	92.6	87.6
	SE-PB-NI-FGSM	**99.9**	**99.3**	**99.9**	**100.0**	**97.5**	**95.3**

The attacking result is also as predicted that we achieve over 95% average success rate, which outperforms the PB methods over 5%.

A.2 ATTACKS WITH SAME FORWARD TIMES

In this part we study another problem that whether the comparsion of four input transformation methods is fair enough. As discussed before in the open-review of SIM, a researcher will doubt whether just setting iteration the same but forwarding and backwarding the model different times for different input transformation methods is fair. We doubt this problems too.

The main idea in this part to study the question is to keep the forward times same for every method and compare the attacking effectiveness instead of keeping the iteration. We know that the forward times for DIM, SIM, VRM and SEM is 1, 5, 20 and 50 during each single iteration. So for keeping the forward times the same when attacking, we adjust the attack iterations of each method as 1000, 200, 50 and 20 separately. The results are shown in Tab.5.

It is noticeable that the transferability of adversarial examaple generated in this way is clearly lower than the results shown in Fig.1. And we thought the biggest problem of causing this problem is overfitting. Although all methods try to enhance the transferability of adversarial examples via various input transformations, the overfitting phenomenon is inevitable when the iteration increases. From the Fig.5, on the one hand we can conclude SEM has higher transferability in this situation as previous setting, and on the other hand, SEM has higher success rate than others in white-box setting too.

A.3 MORE RESULTS OF ATTENTION MAPS

More results of attention maps are illustrated in Fig.4

Table 5: The success rate (%) of attacks with the same forward times.

Model	Attack	Inc-v3	IncRes-v2	Res-50	Den-121	Inc-v3$_{adv}$	IncRes-v2$_{ens}$
Inc-v3	VR-MI-FGSM	99.9*	82.2	68.4	75.0	**60.8**	**43.0**
	DI-MI-FGSM	**100.0***	72.9	59.4	67.4	36.8	19.9
	SI-MI-FGSM	**100.0***	58.5	43.6	56.9	30.0	13.5
	SE-MI-FGSM	**100.0***	**87.4**	**75.9**	**82.2**	49.4	30.1
	VR-NI-FGSM	**100.0***	77.2	62.7	68.9	49.0	28.8
	DI-NI-FGSM	**100.0***	72.0	55.8	64.2	33.3	16.1
	SI-NI-FGSM	**100.0***	52.2	39.5	51.1	26.4	10.7
	SE-NI-FGSM	**100.0***	**90.2**	**77.9**	**83.3**	**52.3**	**30.4**
IncRes-v2	VR-MI-FGSM	84.1	96.6*	73.2	77.2	**64.9**	**61.4**
	DI-MI-FGSM	80.6	99.5*	63.5	69.6	39.8	29.2
	SI-MI-FGSM	81.9	**99.7***	64.0	72.9	52.3	40.1
	SE-MI-FGSM	**88.5**	98.3*	**78.3**	**82.8**	62.2	52.3
	VR-NI-FGSM	84.2	99.7*	68.2	73.8	55.2	43.3
	DI-NI-FGSM	81.7	**100.0***	63.1	70.3	38.3	24.0
	SI-NI-FGSM	76.1	99.9*	54.3	67.0	40.4	28.6
	SE-NI-FGSM	**90.4**	99.5*	**77.5**	**84.0**	**57.3**	**43.4**
Res-50	VR-MI-FGSM	61.5	55.0	98.4*	72.4	36.5	36.0
	DI-MI-FGSM	50.0	41.6	97.8*	55.9	16.5	12.9
	SI-MI-FGSM	39.1	29.1	95.8*	49.9	15.8	11.2
	SE-MI-FGSM	**96.9**	**96.0**	**100.0***	**99.1**	**52.2**	**42.6**
	VR-NI-FGSM	37.9	29.8	91.8*	50.3	20.8	17.2
	DI-NI-FGSM	41.2	33.6	93.2*	48.0	13.8	9.4
	SI-NI-FGSM	27.7	20.5	82.6*	36.6	11.9	7.8
	SE-NI-FGSM	**95.7**	**95.1**	**100.0***	**99.3**	**45.3**	**33.3**
Den-121	VR-MI-FGSM	48.8	40.2	50.5	99.3*	30.2	29.5
	DI-MI-FGSM	42.4	34.6	36.6	99.3*	15.2	12.0
	SI-MI-FGSM	30.8	22.4	26.7	98.5*	13.8	10.4
	SE-MI-FGSM	**92.4**	**89.4**	**94.8**	**100.0***	**49.5**	**37.5**
	VR-NI-FGSM	32.8	25.4	32.8	97.7*	19.3	16.9
	DI-NI-FGSM	39.5	32.1	30.6	98.2*	13.5	9.6
	SI-NI-FGSM	25.8	17.8	20.9	94.4*	12.4	6.8
	SE-NI-FGSM	**93.1**	**88.7**	**95.7**	**100.0***	**45.1**	**28.5**

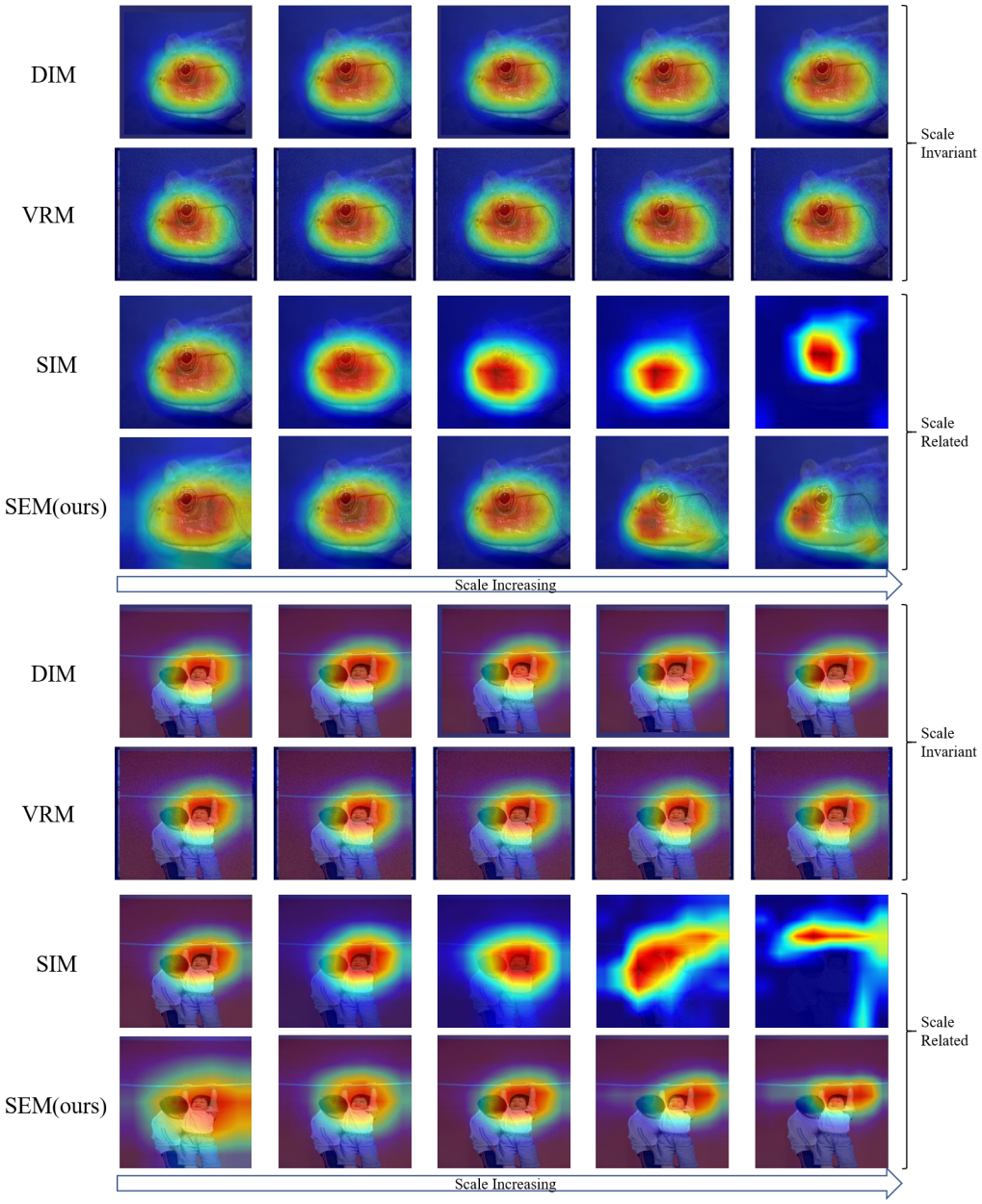


Figure 4: More results of target model's attention maps on the transformed images.