

Transferable adversarial attack based on attention space

Bie Yunlong¹, Sun Jiaze¹(✉), Zhang Yinghui²

1. School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

2. School of Cyberspace Security, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Abstract

Adversarial attacks are a major challenge to the security of deep neural networks (DNNs). However, current transferable attacks often produce perturbations deterministically. This leads to poor local optima, resulting in low transferability. To address this issue, an attention-space transfer attack (ASTA) is proposed to enhance transferability through randomized disruption of salient features. Specifically, a randomized perturbation strategy utilizing latent codes is employed to disrupt shared attention features across models. Furthermore, attention and diversity loss functions are introduced to explicitly guide the generator in exploring the perturbation search space, thereby avoiding overfitting to surrogate models. Experimental results demonstrate that this method achieves an average attack success rate (ASR) of 83.6% on multiple black-box target ensemble models, 6.7% higher than the baseline methods. It effectively generates adversarial samples and demonstrates stronger transferability.

Keywords transferable adversarial attacks, attention space, generative models, diversity regularization, image classification

1 Introduction

Despite the significant performance of DNNs in visual tasks^[1], existing research^[2] has revealed their vulnerability to adversarial examples. Adversarial attacks are classified into white-box and black-box attacks depending on the attacker's knowledge of the model. Recent studies have focused on the more challenging black-box attacks. Query-based attacks^[3] estimate the unknown model's gradient using query outputs, but the excessive number of required queries limits their practicality. In contrast, transfer-based attacks rely on transferability, which enables them to deceive black-box target models, attracting more

attention. In fact, the study of adversarial robustness has extended beyond computer vision to other domains, such as multi-agent systems and wireless communications^[4], highlighting the broad relevance of developing robust defense and attack strategies.

However, traditional attacks^[5] tend to overfit to the surrogate model, leading to poor transferability. To address this issue, improved optimization algorithms have been proposed. Dong et al.^[6] applied momentum to avoid poor local optima. Xie et al.^[7] introduced random transformations to the images. Dong et al.^[8] proposed a translation-invariant attack, and Wang et al.^[9] applied variance tuning for more stable momentum. Based on the observation that different models learn similar features, Zhou et al.^[10] maximized the distance between the features of the original image and the adversarial example. However, classifiers often learn model-specific features^[11], and

recent research has focused more on perturbing salient features; attention mechanisms have demonstrated excellent performance in image recognition tasks with complex backgrounds^[12]. Therefore, adopting this mechanism might have a positive impact on ASTA. Wu et al.^[13] disrupted attention heatmaps, while Wang et al.^[14] aggregated gradients to perturb object-aware features. Zhang et al.^[15] utilized gradient information to evaluate the contribution of each neuron to the final classification decision and selectively perturbed the important neurons based on their contribution magnitude.

However, these attacks rely on gradient-based methods^[5] that generate perturbations in a

deterministic manner. In each iteration, the perturbation is updated in a single specific direction to maximize the given loss function. As illustrated in Fig. 1, this lack of randomness often prevents full exploration of the loss surface, causing adversarial examples to greedily fall into narrow local optima and overfit to the surrogate model. In contrast, introducing high diversity enables a broader exploration of the adversarial subspace. This comprehensive search increases the probability of identifying shared vulnerable features that effectively transfer across different model architectures, rather than those that only fool the specific surrogate model.

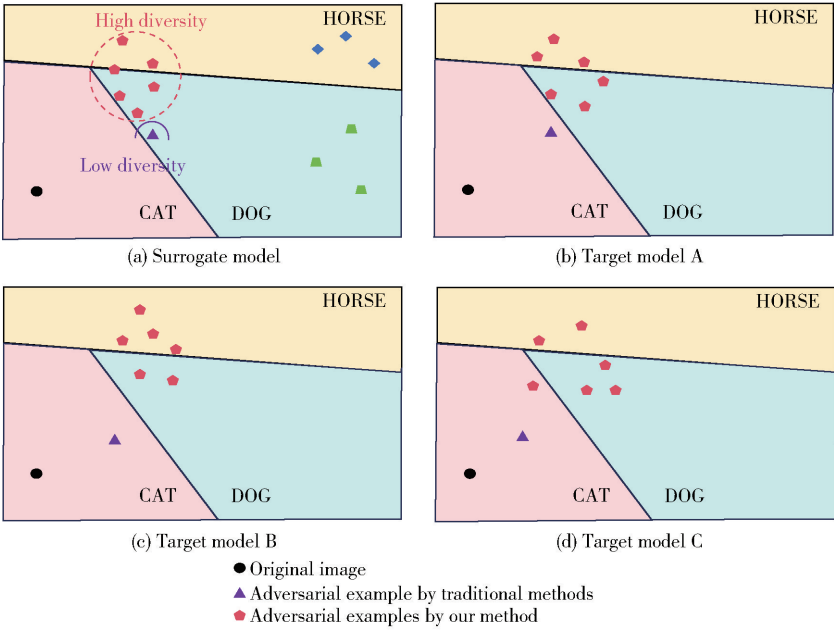


Fig. 1 Adversarial samples on class decision boundaries

To address this issue, ASTA is proposed, which enhances the transferability of adversarial examples by diversely perturbing salient features. Based on recent findings^[12–13], perturbing salient features can improve transferability. This paper perturbs image attention, which highlights the features responsible for the model’s decision and may be shared across different models. However, greedily using gradient-based methods to erode attention can lead to deterministic adversarial examples that easily fall into poor local optima. To avoid this, ASTA perturbs these features in a diverse and random manner. Specifically, the generator is

guided to produce diverse perturbations based on input latent codes, perturbing the attention in different ways. As shown in Fig. 1, this approach enables a more comprehensive exploration of the search space for transferable adversarial examples. The generator learns to produce diverse perturbations that lie outside poor local optima. These adversarial examples effectively deceive target models, while adversarial examples generated by existing deterministic methods tend to overfit the surrogate model.

In summary, the specific contributions of this paper are as follows.

1) This paper introduces randomness at the feature level to comprehensively explore the search space for transferable perturbations, thereby improving the transferability of adversarial examples.

2) A transferable adversarial attack based on attention space is proposed, which perturbs image attention in a diverse, non-deterministic manner.

3) The generator is improved, and attention loss and diversity loss functions are constructed to guide the generation of transferable adversarial examples.

4) Experiments show that ASTA improves ASR across multiple black-box target models and demonstrates better transferability.

2 Related work

2.1 Class activation mapping (CAM)

CAM is a popular technique used to understand how convolutional neural networks (CNNs) perform image recognition and localization. Zhou et al.^[16] proposed CAM to address the weakly supervised learning problem of object localization across different categories. By applying global average pooling (GAP) and a weight matrix to the final convolutional feature maps of a CNN, CAM generates heatmaps that visualize the model's decision-making process, enhancing the interpretability of CNNs. However, CAM has specific requirements for network architecture, which limits its general applicability. Selvaraju et al.^[17] extended the CAM method and proposed gradient-weighted class activation mapping (Grad-CAM). This method can be applied to CNNs with any architecture without modifying the network or retraining, significantly improving model interpretability. Compared to standard class activation mapping, Grad-CAM offers greater generalization. Deng et al.^[18] proposed a spatial perturbation algorithm based on Grad-CAM, which uses Grad-CAM to quickly locate key decision areas in the image, reducing unnecessary interference. However, the algorithm does not consider transferability.

2.2 Transferable adversarial attacks

In adversarial attack research, attacks are generally

categorized into white-box and black-box types. In practical applications, black-box attacks are more realistic since the model's internal information is inaccessible. This study focuses on transfer-based adversarial attacks, which achieve cross-model attacks through the transferability of adversarial examples.

To enhance adversarial transferability, many studies have attempted to improve gradient computation strategies. The momentum iterative method (MIM)^[6] introduces a momentum term to smooth the update path. However, its effectiveness depends on model precision and gradient quality, making it prone to local optima. The gradient relevance attack (GRA)^[19] improves transferability by analyzing neighborhood information and cosine similarity, but its high computational complexity may affect real-time performance and efficiency.

Input transformation-based approaches have also been widely adopted to improve transferability. Jang et al.^[20] and Dong et al.^[21] generate diverse perturbations from a single image, but their diversity is limited to the pixel level and mainly used to enhance adversarial robustness. The diverse input method (DIM)^[7] increases sample diversity through random transformations, but such transformations may introduce unrealistic distortions and semantic inconsistencies. Moreover, DIM perturbs features only implicitly through pixel-level transformations, lacking explicit feature diversification. The patch-wise iterative method (PIM)^[22] mitigates local overfitting via block-wise processing but may disrupt global semantic consistency. The path-augmented method (PAM)^[23] improves transferability by selecting representative augmentation paths and combining them with a greedy search strategy. The DiffAttack method^[24] simulates a reverse diffusion process that gradually generates adversarial perturbations, thereby enhancing cross-model transferability while maintaining imperceptibility.

In studies focusing on intermediate feature representations, the transferable adversarial perturbations (TAP) method^[10] improves cross-architecture transferability by maximizing the distance between clean images and their adversarial counterparts in intermediate feature maps. However, it may overly rely on a specific model architecture when handling

complex hierarchical features. Feature disruption attack (FDA)^[25] and feature importance-aware attack (FIA)^[14] optimize perturbation generation by measuring neuron importance. Yet, since these methods rely on gradient information to evaluate importance, they are susceptible to gradient saturation, leading to inaccurate estimation of neurons' true contribution to the final decision. The neuron attribution attack (NAA)^[15] addresses this limitation by providing a more precise neuron importance assessment, thereby enhancing the cross-model transferability of adversarial examples.

In this work, we propose a randomized perturbation strategy to comprehensively explore the adversarial search space, avoiding local optima and improving cross-model attack performance. Unlike traditional methods, ASTA introduces attention-guided feature-level perturbations that explicitly disrupt diverse semantic features, thereby preventing semantic distortion. Based on the findings of Lin et al.^[26], which demonstrate that attention mechanisms can effectively improve image recognition accuracy, we incorporate attention guidance into adversarial attack design. This theoretical foundation enables ASTA to enhance adversarial transferability across different models while overcoming the semantic inconsistency problem often caused by excessive reliance on input-level transformations.

3 Diversity perturbation attack method in attention space

Let h_ψ be the target classifier. The goal of untargeted adversarial attacks is to create an adversarial example

x^{adv} from image x in class t , causing misclassification on the target classifier (i. e., $h_\psi(x^{\text{adv}}) \neq t$). In the black-box attack scenario, the attacker cannot access the target classifier. Instead, this paper uses an accessible surrogate model f_θ , which shares the same output space as h_ψ but has a different architecture or parameters. Then, an adversarial example with transferability is generated on the surrogate model,

$$\left. \begin{aligned} & \arg \max_{x^{\text{adv}}} L_\theta(x^{\text{adv}}, t) \\ & \text{s. t.} \\ & \|x - x^{\text{adv}}\|_\infty \leq \varepsilon \end{aligned} \right\} \quad (1)$$

where $L_\theta(\cdot, \cdot)$ is the classification loss on the surrogate model f_θ , and ε is the constraint set on the perturbation magnitude. The L_∞ norm is used to constrain the perturbation magnitude. Mathematically, this constraint is defined as $\|x - x^{\text{adv}}\|_\infty = \max_i |x_i - x_i^{\text{adv}}| \leq \varepsilon$, where i represents the index of pixels. This ensures the validity of adversarial examples in the input space by constraining the maximum change across each feature dimension.

Such transfer-based attack methods heavily rely on their transferability. However, existing methods typically optimize Eq. (1) using gradient-based techniques, greedily finding a deterministic solution to maximize $L_\theta(x^{\text{adv}}, t)$. This deterministic nature may lead to overfitting adversarial examples to the surrogate model, reducing the success rate of black-box attacks.

To address this challenge, this paper proposes a transferable adversarial attack based on attention space. As illustrated in Fig. 2, ASTA significantly improves the transferability of adversarial examples by using an attack generator guided by attention perturbation and feature diversification.

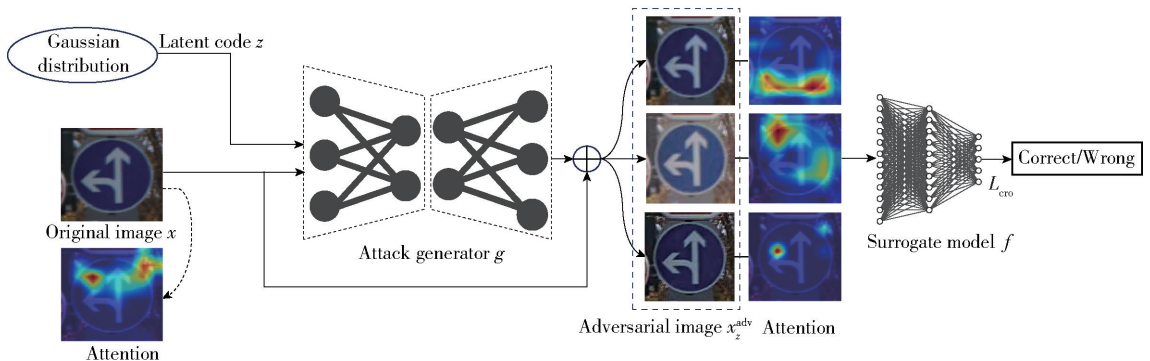


Fig. 2 Workflow of diversity perturbation attack on attention space

Given an input image x and a random latent code z , the image features are first extracted through an encoder composed of multiple convolutional blocks. Then, the attention perturbation module identifies key feature regions in the image, and perturbations are introduced in these regions. Using an attack generator g , combined with the latent code z , diverse adversarial perturbations are generated. The diversity of the perturbations is ensured by calculating the attention loss and diversity loss, preventing the generator from falling into local optima. Finally, the perturbations are added to the original image to generate adversarial examples, and these examples are classified using the surrogate model f . The classification results are used to compute the loss, and gradient information is fed back to the generator to optimize the network parameters. This process enables the generation of more effective adversarial examples, improving the transferability of the attack.

3.1 Attack generator

Unlike gradient-based iterative and deterministic methods for generating perturbations, this paper uses a generator that parameterizes adversarial perturbations through a DNN architecture. Given an image x and a latent code z sampled from a Gaussian distribution, the generator g learns to output adversarial perturbations that depend on the latent code. The adversarial example x_z^{adv} is constructed as

$$x_z^{\text{adv}} = C_{x,\varepsilon} \{ x + \varepsilon \cdot g(x, z) \} \quad (2)$$

where $C_{x,\varepsilon}$ clips the perturbation for each pixel, restricting it within an L_∞ -norm ball of radius ε . This ensures that the magnitude of each pixel's perturbation does not exceed the specified threshold ε . By doing so, it guarantees that the generated adversarial example does not stray too far from the original image, thus maintaining the effectiveness of the adversarial attack.

3.2 Attention perturbation

To enhance transferability, this paper introduces perturbations to the attention areas of the image. The attention regions highlight features critical to the model's decision-making, which are typically related to the main objects in the image. Since different

classifiers usually rely on these object-related features for decision-making, perturbing these features can effectively transfer to other models.

Based on Grad-CAM^[17], the attention A is defined as the weighted representation of features F , which are obtained from the output of the last convolutional layer. For the Inception-v3 (Inc-v3) model, the last convolutional layer is Mixed 7c. The attention A is defined as shown in Eq. (3):

$$A(x; t) = \alpha_t F = G \left(\frac{\partial y_t}{\partial F} \right) F \quad (3)$$

The weight α_t represents the importance of feature F given the true class t . It is obtained by calculating the gradient of the prediction y_t for class t with respect to feature F , followed by applying global average pooling $G(\cdot)$ over the spatial dimensions. To prevent the generator from only perturbing a few channels with the highest magnitudes, this paper further normalizes the attention $A(x; t)$ across channels. Since the feature extraction process involving convolution and activation functions (e. g., rectified linear unit (ReLU)) is fully differentiable, the resulting attention map $A(x; t)$ is also differentiable with respect to the input image. This property is crucial as it allows the entire generator framework to be optimized end-to-end using standard backpropagation.

Then, we treat the attention maps as probability distributions and employ the Wasserstein-1 distance (earth mover's distance) to measure their discrepancy. Formally, the distance between the distribution of the original image P and the adversarial sample Q is defined as $W(P, Q) = \inf_{\gamma \in \Pi(P, Q)} \mathbb{E}_{(u,v) \sim \gamma} [\|u - v\|]$. Based on this metric, the generator learns to maximize the divergence in attention distribution between the adversarial sample and the original image by maximizing the following attention loss L_{attn} ,

$$L_{\text{attn}} = W(A(x_z^{\text{adv}}; t), A(x; t)) \quad (4)$$

Although Wu et al.^[13] also perturbed attention heatmaps extracted using the Grad-CAM^[17] technique, the difference in this paper's approach lies in the additional application of channel normalization and the introduction of a method based on the Wasserstein distance metric. Without channel normalization, the generator would only perturb a few feature channels

with the highest magnitudes, thus limiting the diversity of the generated perturbations. To avoid this, this paper normalizes each feature channel, enabling the generator to perturb a more diverse set of features.

3.3 Feature diversification

In the absence of any guidance, the generator may neglect to learn the latent encoding of the input and greedily maximize the attention loss deterministically, similar to gradient-based methods. Therefore, by introducing the attention loss, the generator is guided to explore and perturb diverse features in a random manner. This paper trains the generator to disturb the attention representation in different ways using two distinct input latent encoding z_1 and z_2 (both sampled from a Gaussian distribution). The cosine similarity is constructed to measure the attention space similarity of the two samples, with the aim of maximizing the diversity loss L_{div} ,

$$L_{\text{div}} = \frac{1 - \cos(A(x_{z_1}^{\text{adv}}; t), A(x_{z_2}^{\text{adv}}; t))}{\|z_1 - z_2\|} \quad (5)$$

By passing z_1 and z_2 into the generator as Eq. (2), two adversarial samples $x_{z_1}^{\text{adv}}$ and $x_{z_2}^{\text{adv}}$ are generated, along with their respective attention representations $A(x_{z_1}^{\text{adv}}; t)$ and $A(x_{z_2}^{\text{adv}}; t)$ (Eq. (3)). Then, by maximizing the distance between these two representations, the generator is compelled to produce semantically diverse perturbations.

Although Yang et al. [27] initially proposed diversity regularization, their application was limited to pixel or feature levels. However, diversity at these levels does not necessarily translate to diversity in the attention space, which may not effectively guide the generator to perturb significant features in a diverse manner. To explicitly guide the generator to perturb meaningful features diversely, this method differs from existing approaches by applying diversity regularization at the attention level.

Overall, the attack generator g is trained by maximizing the following objective,

$$L = L_{\text{cro}} + \lambda_{\text{attn}} L_{\text{attn}} + \lambda_{\text{div}} L_{\text{div}} \quad (6)$$

where L_{cro} is the cross-entropy loss between the adversarial samples and the true labels, and λ_{attn} and λ_{div} control the weights of the attention loss L_{attn} and

diversity loss L_{div} , respectively. The definition of the cross-entropy loss L_{cro} is given as

$$L_{\text{cro}} = - \sum_{c=1}^C y_c \ln p_c \quad (7)$$

where C is the total number of classes, y_c is the true label; if the true class is c , then $y_c = 1$; otherwise, $y_c = 0$. p_c is the model's predicted probability for class c . This loss function is used to measure the difference between the adversarial samples and the true labels.

4 Experiments

4.1 Experimental setup

For the attack generator, we adopt an Attention U-Net architecture, which enhances the traditional U-Net's convolutional encoder-decoder structure [28] through an attention mechanism. It consists of three encoding blocks and three decoding blocks. Each encoding and decoding block is composed of a convolutional layer and a transposed convolutional layer, followed by a batch normalization layer and an ReLU activation layer. Attention modules are integrated into each encoding and decoding block to focus on the key regions of the input, allowing the network to dynamically adjust its focus. In the Attention U-Net, the latent code z is first processed through the attention module to weight the importance of different features, then spatially expanded and concatenated with the input of the block. The generator is trained for 100 epochs with a learning rate of 10^{-4} and a batch size of 8, using the Adam optimizer [29], with hyperparameters set to $\beta_1 = 0.5$ and $\beta_2 = 1$, and weight decay of 10^{-5} .

The paper randomly selects 10 000 images from the ImageNet validation set [30] as training data and 1 000 images from the NeurIPS 2017 adversarial competition [31] as test data. The testing methods are evaluated on the following models: Inception-v3 (Inc-v3) [1], Inception-v4 (Inc-v4) [32], Inception-ResNet-V2 (IncRes-v2) [32], ResNet-V2 (Res-v2) [33], and VGG16 (Vgg-16) [34]. ASTA is compared with various state-of-the-art attack methods, including momentum iterative fast gradient sign method (MI-FGSM) [6],

DIM^[7], variance-tuning momentum interative fast gradient sign method (VMI-FGSM)^[9], TAP^[10], FIA^[14], PAM^[23], NAA^[15], and learning to transform (L2T)^[35]. The number of iterations for these methods is set to $T = 10$, with a step size of $\alpha = 1.6$, while other hyperparameters follow the settings in their respective references. The maximum perturbation constraint is set to 16 under the L_∞ norm. For ASTA, the weight parameters are set to $\lambda_{\text{attn}} = 10$ and $\lambda_{\text{div}} = 1\,000$, and the length of the latent encoding z is set to 16. Note that while ASTA involves a one-time training cost, it enables real-time attack generation via a single forward pass, offering superior inference efficiency

compared to iterative baselines.

4.2 Transferability comparison

To compare the transferability of ASTA with baseline methods, adversarial samples were generated on three surrogate models: Inc-v3, Inc-v4, and IncRes-v2. The ASR, defined as the error classification rate of the models, were measured on five target models, with the results presented in Table 1. The ensemble model (Ensemble) represents the ASR of all target models excluding the white-box model, with predictions based on the average probability output from all models.

Table 1 ASR of different methods on various target models %

Surrogate model	Attack	Target model					
		Inc-v3	Inc-v4	IncRes-v2	Res-v2	Vgg-16	Ensemble
Inc-v3	MI-FGSM	97.8	42.7	40.1	41.2	53.2	35.9
	DIM	98.1	68.5	61.8	53.3	68.7	58.2
	VMI-FGSM	97.7	69.5	66.9	57.6	70.1	61.7
	TAP	99.7	77.9	75.3	53.1	70.5	69.1
	FIA	98.6	84.2	80.2	69.3	85.6	77.6
	PAM	99.8	83.7	81.2	77.5	81.3	80.8
	NAA	98.1	85.3	82.4	77.1	87.6	81.5
	L2T	99.6	88.2	85.7	81.6	93.8	85.0
	ASTA	96.3	88.7	83.2	82.4	95.4	85.3
Inc-v4	MI-FGSM	59.7	98.7	44.9	47.7	63.6	44.8
	DIM	75.3	98.1	66.6	60.5	74.8	66.5
	VMI-FGSM	76.6	98.4	70.5	65.2	76.9	68.7
	TAP	75.4	99.7	70.2	59.6	82.6	72.2
	FIA	83.2	95.3	78.4	74.7	85.3	78.3
	PAM	87.5	99.6	83.5	77.6	85.4	82.9
	NAA	86.1	96.5	81.3	75.5	83.7	80.8
	L2T	92.3	99.6	85.2	78.8	89.5	83.7
	ASTA	85.5	97.7	67.6	79	89.6	79.5
IncRes-v2	MI-FGSM	58.1	52.6	99.4	46.9	63.5	47.5
	DIM	73.2	70.5	94.9	57.7	70.4	65.8
	VMI-FGSM	78.4	77.4	99.4	64.5	74.2	71.6
	TAP	74.1	66.8	95.2	46.9	63.8	51.9
	FIA	81.9	77.1	88.2	66.9	81.9	74.8
	PAM	86.8	88.4	99.3	80.3	86.1	85.2
	NAA	82.4	78.3	93.2	74.4	83.6	78.3
	L2T	90.6	88.5	99.4	85.8	89.4	84.7
	ASTA	82.5	89.5	93.1	80.8	89.5	85.9

The results show that ASTA achieves higher ASR on most black-box target models compared with existing approaches, with an average improvement of 7.6% over FIA on the ensemble model. Compared with PAM and NAA, ASTA demonstrates significant advantages across different target models: on Inc-v3, the ASR increases by 4.5% over PAM and 3.8% over NAA; on Inc-v4, the ASR improves by 0.7% over NAA; and on Inc-Res-v2, the ASR rises by 7.6% over NAA. The L2T method performs well on some target models (e. g., Inc-v4 and Inc-Res-v2), but its overall transferability on the ensemble model is slightly lower than that of ASTA. This indicates that although L2T has advantages on specific architectures, its generalization ability is less stable than ASTA. In addition, although ASTA achieves slightly lower ASR on white-box target models, it shows a significant improvement on black-box targets, suggesting reduced overfitting to the surrogate model and better generalization to unseen models.

It can be observed from Table 1 that when the architectural difference between the target model and the surrogate model is greater, ASTA outperforms existing methods. For instance, the adversarial samples generated by the proposed attack method on Inc-v3 achieve only a 4.5% higher ASR than FIA on the structurally similar Inc-v4^[32]; however, they outperform by 9.8% on the structurally different Vgg-16^[34]. Existing methods often overfit to the surrogate models, showing lower transferability on models with significant structural differences, whereas ASTA demonstrates high transferability across different model architectures.

4.3 Diversified perturbation generation

Recent studies^[36–37] have shown that a model’s adversarial robustness improves when it is trained on diversified adversarial samples. Based on this idea, the robustness of classifiers trained using the generated adversarial samples was tested to demonstrate that ASTA can indeed generate diversified perturbations. The Inc-v3 model was trained on the Caltech101 dataset^[38], which

contains 8 681 images across 101 categories, randomly split into training and testing sets of 7 332 and 1 349 images, respectively. The training lasted for 50 epochs, using a batch size of 8, with the stochastic gradient descent (SGD) optimizer set to a learning rate of 0.001, momentum of 0.9, and weight decay of 5×10^{-4} . In each epoch, adversarial examples were generated on-the-fly using the designated attack strategy, and the model was updated to minimize the cross-entropy loss on these examples.

Table 2 presents the classification accuracy of the classifier under various attacks using different adversarial training methods. The adversarial samples used for evaluation were generated on classifiers trained with the original images (no defense). In all cases, the classifiers using ASTA with L_{div} demonstrated the highest robustness, averaging 2.98% higher than the classifiers that did not use L_{div} . While the methods not employing L_{div} and the gradient-based methods (FGSM^[2], BIM^[5], and PGD^[37]) trained the classifiers only against a limited set of deterministic attacks, the method using L_{div} was able to train classifiers with diversified adversarial samples, thereby enhancing the robustness of the classifiers.

Table 2 Classification accuracy of adversarially trained models under different attacks

Method	%				
	No attack	FGSM	BIM	PGD	ASTA
No defence	91.86	35.53	2.88	2.62	4.46
FGSM	79.26	82.26	81.12	84.03	83.77
BIM	83.17	82.88	83.19	83.14	82.28
PGD	85.83	84.73	85.63	85.55	85.15
ASTA (without L_{div})	89.17	80.81	85.63	86.66	86.93
ASTA	90.89	85.93	88.32	89.27	89.65

Additionally, Fig. 3 visualizes the semantically diverse adversarial samples generated by ASTA. Starting from a single image, this method can generate multiple adversarial samples that perturb the image’s attention and the model’s predictions in diverse ways.

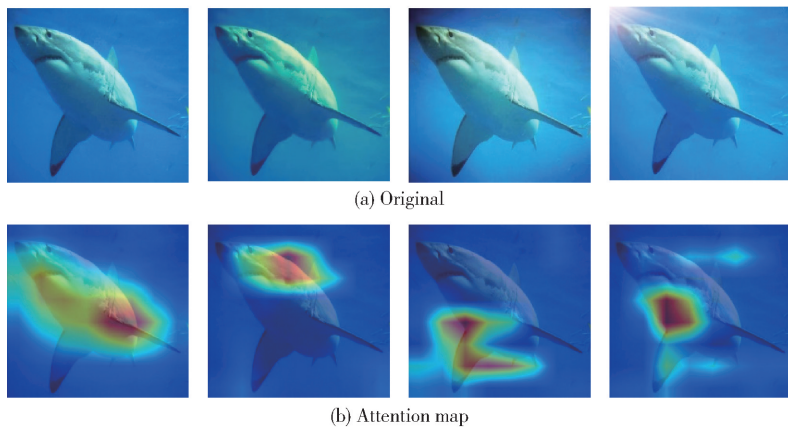


Fig. 3 Original samples and their attention maps

4.4 The impact of diversity on transferability

To demonstrate that perturbing diverse features can improve transferability, Fig. 4 visualizes the features of adversarial samples generated by both FIA and ASTA, projecting them into a two-dimensional space formed by feature vectors obtained through principle component analysis (PCA).

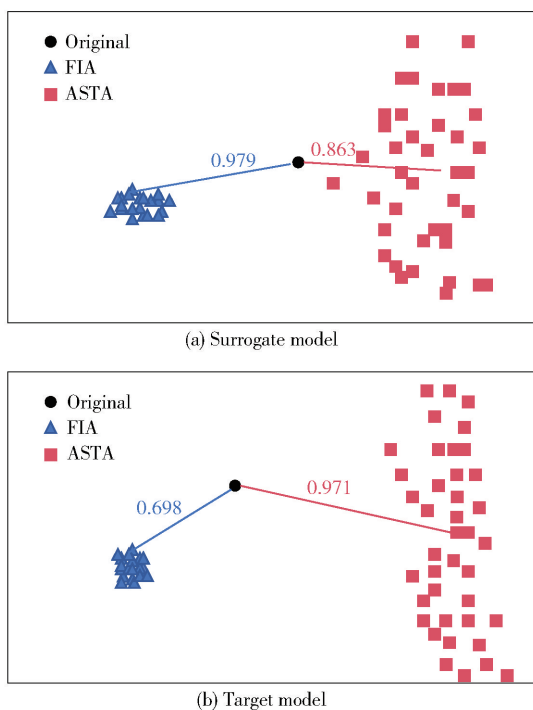


Fig. 4 Features of adversarial samples generated by different methods

The PCA visualization of features from adversarial samples generated by FIA and ASTA on the surrogate

model (Inc-v3) and the target model (Res-v2) reveals that the feature representations of adversarial samples generated by FIA form a dense cluster, while those generated by ASTA are more widely distributed. Additionally, the distribution shown in Fig. 4 (a) indicates that adversarial samples generated by FIA are stronger on the surrogate model compared to those generated by ASTA; the average L_2 distance between the features of adversarial samples generated by FIA and the original image features is 0.979, significantly higher than the 0.863 of ASTA. However, the adversarial samples generated by FIA show poor transferability on the target model, with an average distance of 0.698, far lower than the 0.971 of ASTA. By learning to generate semantically diverse perturbations, ASTA effectively avoids poor local optima and enhances transferability.

4.5 Parameter influence

4.5.1 Different disturbance constraints

This paper also compares the transferability of the proposed attack method with baseline methods under lower disturbance constraints ε (i. e. , 4, 8, and 12). As shown in Fig. 5, the ASR on the black-box target model ensemble is presented based on Inc-v3 as the surrogate model. Under all constraints, ASTA outperforms existing methods in terms of transferability, averaging 8.2% higher than FIA. Even when the disturbances are less noticeable, this method still demonstrates high transferability.

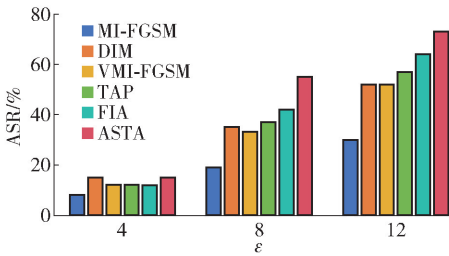


Fig. 5 ASR on the black-box target model ensemble under different perturbation constraint values

4.5.2 Impact of weight

Fig. 6 shows the influence of the weight λ_{attn} on the ASR when $\lambda_{\text{div}} = 1\ 000$ based on Inc-v3 as the surrogate model. This weight is used to control the attention loss L_{attn} . The weight L_{attn} plays a significant role in enhancing transferability, achieving the highest performance when $\lambda_{\text{attn}} = 10$. Using a slightly lower weight of $\lambda_{\text{attn}} = 1$ increases the ASR on Inc-v4, but it reduces the ASR on other black-box models. Insufficient weighting on the attention loss can lead to overfitting of the attack to the surrogate model, improving transferability on the structurally similar Inc-v4 while decreasing it on black-box models with greater structural differences. Excessive focus on L_{attn} can interfere with the attack’s disruption of the classifier’s predicted output, resulting in decreased ASR across all models.

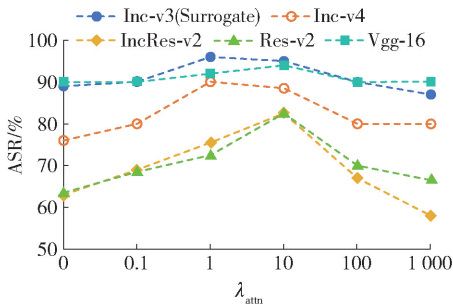


Fig. 6 Impact of different attention loss weights on ASR

4.5.3 Impact of diversity loss weight

Fig. 7 shows the influence of different diversity loss weights λ_{div} on the ASR when $\lambda_{\text{attn}} = 10$ based on Inc-v3 as the surrogate model. Using L_{div} generally improves transferability, reaching its peak performance when $\lambda_{\text{div}} = 1\ 000$, indicating that diversifying perturbation features can prevent adversarial samples

from overfitting to the surrogate model. At the same time, in this case, using a lower $\lambda_{\text{div}} = 100$ increased the ASR on Inc-v4 but significantly harmed the ASR on other target models, demonstrating that a lack of diversity led the generator to overfit to the structurally similar Inc-v3 and Inc-v4. Excessively high diversity loss weights can hinder the generator from perturbing features in a destructive manner, thereby reducing the ASR.

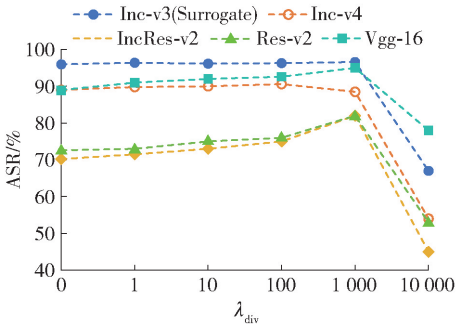


Fig. 7 Impact of different diversity loss weights on ASR

4.6 Ablation study

An ablation study was conducted to validate the effectiveness of various components in ASTA. In all experiments, IncRes-v2^[21] was used as the surrogate model to generate adversarial samples.

In Sect. 3.2, channel normalization was applied to the attention loss L_{attn} to prevent perturbations from being concentrated on only a few feature channels with high amplitudes. To verify this claim, the normalization process was removed from the framework, and its transferability was tested. The results are shown in Table 3, indicating that the method utilizing channel normalization achieved higher ASR on all models compared to the method that did not use channel normalization.

Table 3 Impact of attention loss L_{attn} with channel normalization on ASR

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-v2	Vgg16	Ensemble
ASTA (without channel-norm)	79.4	87.3	93.2	72.1	88.7	82.9
ASTA	85.3	89.6	93.5	80.2	89.6	86.2

This suggests that channel normalization plays an important role in diversifying the perturbation of the image attention.

In Sect. 3.3, by applying diversity at the attention level, the attack generator is explicitly guided to disrupt significant features in a diversified manner. To verify this claim, the diversity L_{div} was modified to maximize the distance of the generated adversarial samples at the pixel or feature level, and their ASR were tested. The results are shown in Table 4, indicating that applying diversity at the feature level achieves higher transferability compared to applying diversity at the pixel level, as it guides the generator to perturb image features in a diversified manner. However, simply perturbing diverse features does not necessarily lead to diversity in the attention space; applying diversity in the attention space yields higher transferability.

Table 4 Impact of applying diversity loss L_{div} at different levels on ASR

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-v2	Vgg16	Ensemble
ASTA	85.3	89.5	93.1	80.9	89.3	85.8
Feature-level	76.4	85.3	92.5	76.4	88.7	82.6
Pixel-level	76.2	79.1	91.8	73.7	88.3	74.7

5 Conclusions

In this study, a transferable adversarial attack based on attention space is proposed, which significantly enhances the transferability of adversarial samples by introducing diversified perturbations within the attention space. The core idea is to leverage the attention mechanism to identify key features in images and rely on the generator to produce perturbations in a non-deterministic manner, avoiding poor local optima and reducing overfitting to the surrogate model. This disrupts significant features in the images that are often shared by multiple models, allowing adversarial samples to effectively attack unknown models. The method has been experimentally validated on multiple datasets and models, demonstrating its ability to generate effective adversarial samples on target models while exhibiting stronger transferability. Future work will aim to improve this attack method and explore its application in zero-shot attacks, further enhancing the security and reliability of models.

Acknowledgements

This work was supported by the National Natural Science

Foundation of China (62272387), the Key Science and Technology Project of Shaanxi Provincial Department of Education (25JS118).

References

- [1] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), 2016, Jun 27 – 30, Las Vegas, NV, USA. Piscataway, NJ, USA: IEEE, 2016; 2818 – 2826.
- [2] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples. Proceedings of the 3rd International Conference on Learning Representations (ICLR'15), 2015, May 7 – 9, San Diego, CA, USA. 2015.
- [3] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information. Proceedings of the 35th International Conference on Machine Learning (PMLR'80), 2018, Jul 10 – 15, Stockholm, Sweden. 2018; 2137 – 2146.
- [4] LÜ Z F, XIAO L, CHEN Y F, et al. Safe multi-agent reinforcement learning for wireless applications against adversarial communications. IEEE Transactions on Information Forensics and Security, 2024, 19: 6824 – 6839.
- [5] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world. Proceedings of the 5th International Conference on Learning Representations (ICLR'17), 2017, Apr 24 – 26, Toulon, France. 2017; 14p
- [6] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18), 2018, Jun 18 – 23, Salt Lake City, UT, USA. Piscataway, NJ, USA: IEEE, 2018; 9185 – 9193.
- [7] XIE C, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19), 2019, Jun 15 – 20, Long Beach, CA, USA. Piscataway, NJ, USA: IEEE, 2019; 2730 – 2739.
- [8] DONG Y P, PANG T Y, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19), 2019, Jun 15 – 20, Long Beach, CA, USA. Piscataway, NJ, USA: IEEE, 2019; 4312 – 4321.
- [9] WANG X S, HE K. Enhancing the transferability of adversarial attacks through variance tuning. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21), 2021, Jun 20 – 25, Nashville, TN, USA. Piscataway, NJ, USA: IEEE, 2021; 1924 – 1933.
- [10] ZHOU W, HOU X, CHEN Y J, et al. Transferable adversarial perturbations. Proceedings of the 15th European Conference on Computer Vision (ECCV'18), 2018, Sept 8 – 14, Munich, Germany. Berlin, Germany: Springer, 2018; 471 – 486.
- [11] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features. Advances in Neural Information Processing Systems 32: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS'19), 2019, Dec 8 – 14, Vancouver, Canada. 2019; 12p
- [12] ZHANG W D, CHEN G C, ZHUANG P X, et al. CATNet: Cascaded attention transformer network for marine species image

- classification. *Expert Systems with Applications*, 2024, 256: Article 124932.
- [13] WU W B, SU Y X, CHEN X X, et al. Boosting the transferability of adversarial samples via attention. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*, 2020, Jun 13 – 19, Seattle, WA, USA. Piscataway, NJ, USA: IEEE, 2020: 1161 – 1170.
 - [14] WANG Z B, GUO H C, ZHANG Z F, et al. Feature importance-aware transferable adversarial attacks. *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV'21)*, 2021, Oct 10 – 17, Montreal, Canada. Piscataway, NJ, USA: IEEE, 2021: 7639 – 7648.
 - [15] ZHANG J P, WU W B, HUANG J T, et al. Improving adversarial transferability via neuron attribution-based attacks. *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, 2022, Jun 18 – 24, New Orleans, LA, USA. Piscataway, NJ, USA: IEEE, 2022: 14993 – 15002.
 - [16] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 2016, Jun 27 – 30, Las Vegas, NV, USA. Piscataway, NJ, USA: IEEE, 2016: 2921 – 2929.
 - [17] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the 2017 IEEE/CVF International Conference on Computer Vision (ICCV'17)*, 2017, Oct 22 – 29, Venice, Italy. Piscataway, NJ, USA: IEEE, 2017: 618 – 626.
 - [18] DENG T, ZENG Z G. Generate adversarial examples by spatially perturbing on the meaningful area. *Pattern Recognition Letters*, 2019, 125: 632 – 638.
 - [19] ZHU H G, REN Y C, SUI X Y, et al. Boosting adversarial transferability via gradient relevance attack. *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV'23)*, 2023, Oct 1 – 6, Paris, France. Piscataway, NJ, USA: IEEE, 2023: 4741 – 4750.
 - [20] JANG Y, ZHAO T C, HONG S, et al. Adversarial defense via learning to generate diverse attacks. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV'19)*, 2019, Oct 27 – Nov 2, Seoul, Republic of Korea. 2019: 2740 – 2749.
 - [21] DONG Y P, DENG Z J, PANG T Y, et al. Adversarial distributional training for robust deep learning. *Advances in Neural Information Processing Systems 33: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS'20)*, 2020, Dec 6 – 12, Vancouver, Canada. 2020: 8270 – 8283.
 - [22] GAO L L, ZHANG Q L, SONG J K, et al. Patch-wise attack for fooling deep neural network. *Proceedings of the 16th European Conference on Computer Vision (ECCV'20)*, 2020, Aug 23 – 28, Glasgow, UK. Berlin, Germany: Springer, 2020: 307 – 322.
 - [23] ZHANG J P, HUANG J T, WANG W X, et al. Improving the transferability of adversarial samples by path-augmented method. *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, 2023, Jun 17 – 24, Vancouver, Canada. Piscataway, NJ, USA: IEEE, 2023: 8173 – 8182.
 - [24] CHEN J Q, CHEN H, CHEN K Y, et al. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, 47(2): 961 – 977.
 - [25] GANESHAN A, BS V, BABU R V. FDA: Feature disruptive attack. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV'19)*, 2019, Oct 27 – Nov 2, Seoul, Republic of Korea. 2019: 8069 – 8079.
 - [26] LIN C X, HU X J, ZHAN Y F, et al. MobileNetV2 with spatial attention module for traffic congestion recognition in surveillance images. *Expert Systems with Applications*, 2024, 255 (D): Article 124701.
 - [27] YANG D D, HONG S H, JANG Y S, et al. Diversity-sensitive conditional generative adversarial networks. *Proceedings of the 7th International Conference on Learning Representations (ICLR'19)*, 2019, May 6 – 9, New Orleans, LA, USA. 2019: DOI:10.48550/arXiv.1901.09024.
 - [28] POURSAEED O, KATSMAN I, GAO B C, et al. Generative adversarial perturbations. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)*, 2018, Jun 18 – 23, Salt Lake City, UT, USA. Piscataway, NJ, USA: IEEE, 2018: 4422 – 4431.
 - [29] KINGMA D P, BA J L. ADAM: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, 2015, May 7 – 9, San Diego, CA, USA. 2015: DOI:10.48550/arXiv.1412.6980.
 - [30] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, 115(3): 211 – 252.
 - [31] KURAKIN A, GOODFELLOW I, BENGIO S, et al. Adversarial attacks and defences competition. *The NIPS'17 Competition: Building Intelligent Systems*. Berlin, Germany: Springer, 2018: 195 – 231.
 - [32] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, 31(1): 4278 – 4284.
 - [33] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks. *Proceedings of the 14th European Conference on Computer Vision (ECCV'16)*, 2016, Oct 11 – 14, Amsterdam, The Netherlands. Berlin, Germany: Springer, 2016: 630 – 645.
 - [34] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, 2015, May 7 – 9, San Diego, CA, USA. 2015: DOI:10.48550/arXiv.1409.1556.
 - [35] ZHU R Y, ZHANG Z L, LIU Z, et al. Learning to transform dynamically for better adversarial transferability. *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, 2024, Jun 16 – 22, Seattle, WA, USA. Piscataway, NJ, USA: IEEE, 2024: 24273 – 24283.
 - [36] SU D, ZHANG H, CHEN H G, et al. Is robustness the cost of accuracy: A comprehensive study on the robustness of 18 deep image classification models. *Proceedings of the 15th European Conference on Computer Vision (ECCV'18)*, 2018, Sept 8 – 14, Munich, Germany. Berlin, Germany: Springer, 2018: 631 – 648.
 - [37] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks. *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, 2018, Apr 3 – May 3, Vancouver, Canada. 2018: 4138 – 4160.
 - [38] LI F F, FERGUS R, PERONA P. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, 28(4): 594 – 611.