SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

June 2026, Vol. 69, Iss. 6, 162105:1–162105:14
https://doi.org/10.1007/s11432-024-4880-x

MAD: meta adversarial defense benchmark

Xiaoxu PENG, Dong ZHOU*, Guanghui SUN, Jiaqi SHI & Ligang WU

Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, China

Received 16 June 2024/Revised 21 December 2024/Accepted 9 October 2025/Published online 23 April 2026

Abstract Adversarial training (AT) is a fundamental technique employed to defend against adversarial attacks and effectively enhance model robustness. In particular, rapid adaptation to unknown attacks with high accuracy is critical for sophisticated and responsive systems, such as autonomous driving systems. Therefore, to address these issues, we propose a novel meta adversarial defense (MAD) benchmark. This benchmark consists of three MAD datasets generated from 30 types of attacks on MNIST, CIFAR-10, and Tiny-ImageNet datasets, along with an evaluation toolkit. In addition, we introduce a meta-learning-based AT (Meta-AT) algorithm as the baseline, with high robustness to unknown adversarial attacks through few-shot learning. Experimental results demonstrate the effectiveness of our Meta-AT compared to the state-of-the-art (SOTA) approaches, such as traditional AT, Fast-AT, Free-AT, adversarial training with transferable adversarial examples (ATTA), and you only propagate once (YOPO). Moreover, the models trained with Meta-AT maintain excellent standard classification accuracy on clean examples (SA) and robust classification accuracy on adversarial examples (RA). This benchmark demonstrates significant improvements in investigating the transferability of adversarial defense methods to unknown attacks and the capacity to learn from a limited number of adversarial examples. Our code and the attacked datasets will be available at https://github.com/PXX1110/MAD.

Keywords adversarial training, adversarial attack, meta adversarial defense, meta-learning, few-shot learning

Citation Peng X X, Zhou D, Sun G H, et al. MAD: meta adversarial defense benchmark. Sci China Inf Sci, 2026, 69(6): 162105, https://doi.org/10.1007/s11432-024-4880-x

1 Introduction

Adversarial examples are an important and interesting topic in deep learning, generally referring to the fact that adding specific small perturbations to a clean example can fool a well-trained model and sometimes even affect the physical world. This poses a threat to cutting-edge applications such as autonomous driving [1], facial recognition [2], and intelligent healthcare [3]. As a result, researchers have focused on exploring different defense schemes to protect against such adversarial attacks, which have received significant attention. Existing adversarial defense methods can be broadly categorized into several types: model alteration [4], detection as defense [5,6], input transformation [7,8], certified defenses [9,10], and other defenses [11,12]. AT [13] is a widely used framework in model alteration. It is acknowledged as one of the strongest principled defenses against adversarial attacks and has inspired the development of many other adversarial defense methods.

Recently, researchers have increasingly focused on leveraging AT to enhance model robustness for defensive purposes. Building on the Bag of Tricks [14], many comprehensive robust benchmarks have emerged. Robust Principle [15], RobustART [16], and ARES-Bench [17] address model structures and training techniques against large adversarial and out-of-distribution (OOD) data. They are more committed to the diversity of data. However, few studies have considered defense against diverse attacks. Beyond the traditional classification of black-box and white-box attacks [18], it can also be categorized into online and offline adversarial attacks based on their execution methods. In traditional AT tasks, real-time attacks like the PGD attack are considered online attacks, and the corresponding defense methods are termed online AT. In contrast, offline adversarial attacks are pre-generated, such as real-world physical attacks generated based on PGD. Effective defense methods against such attacks often rely on offline AT [4]. Preliminary experiments indicate that adversarial examples (AEs) generated offline also impact models trained with online AT. As shown in Figure 1, the performance of five SOTA robust AT methods against offline adversarial attacks on MAD-C is significantly poor. Through extensive research and development, we have identified three open issues in the existing AT-based approaches from a more holistic perspective.

Q1: Challenge of generalized adversarial defenses. It has been found that the defense effectiveness of AT is highly coupled with adversarial attacks. That is, the trained model tends to exhibit strong defense capabilities

* Corresponding author (email: dongzhou@hit.edu.cn)

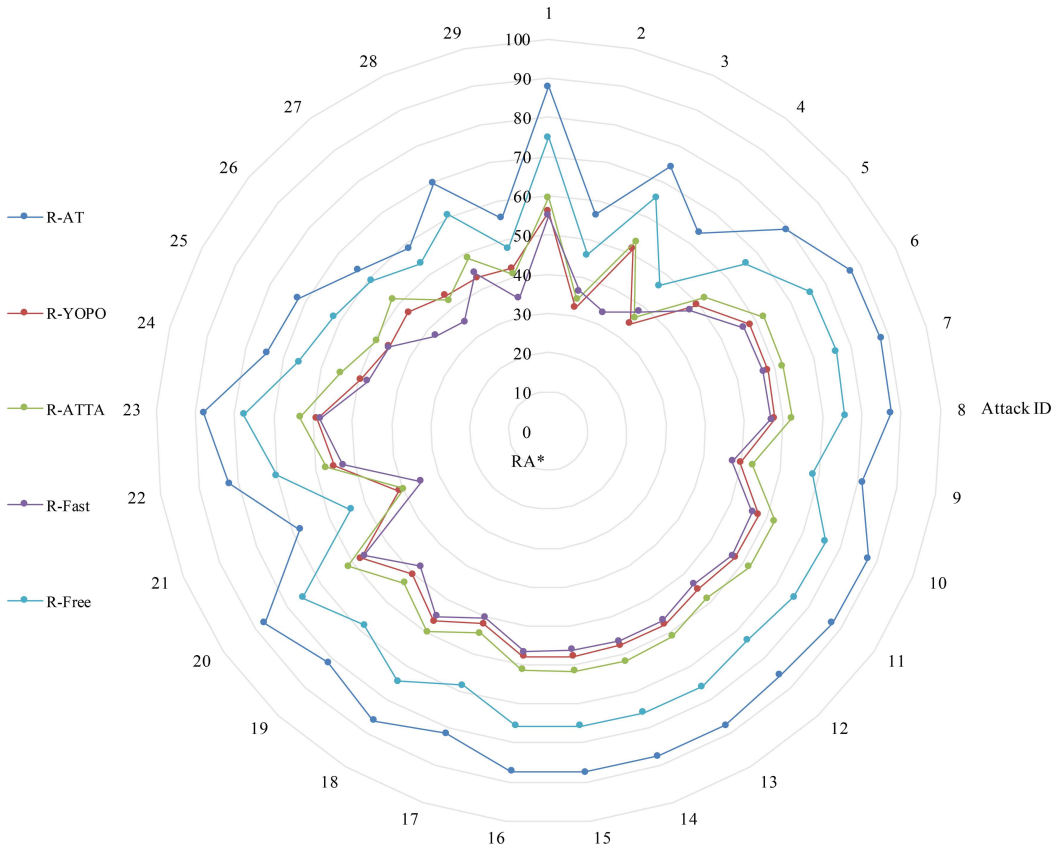


Figure 1 (Color online) Robust classification accuracy following adversarial defense (RA^*, %) of five AT methods against various offline adversarial attacks.

against the attacks used in the training process, and many defense methods are helpless without knowledge of the attack strategy. Developing adversarial defense methods that can withstand new attack techniques remains an ongoing challenge.

Q2: Constraint of online AT. Online AT generates and uses AEs in real-time during training, focusing primarily on gradient-based attacks to enhance defense capability and improve model robustness. In contrast, offline AT pre-generates AEs and fine-tunes the model to defend against prolonged black-box and patch attacks. Existing benchmarks have mostly focused on online AT, neglecting the defense against offline attacks.

Q3: Vulnerability of cross-model adversarial. Cross-model adversarial vulnerability refers to the susceptibility of models to AEs generated by different models. On the one hand, even robustly trained models can fall prey to AEs generated by vanilla models trained on clean data. On the other hand, this is also a common form of some black-box attacks and real-world physical attacks.

To address these issues, we propose an MAD benchmark, which combines the advantages of both online and offline AT to create a robust defense model and support continuous learning. This benchmark includes three MAD-M, MAD-C, and MAD-T datasets, along with an MAD evaluation toolkit. The MAD datasets are constructed by 30 mainstream adversarial attacks on the MNIST [19], CIFAR-10 [20], and Tiny-ImageNet [21]. In the MAD evaluation toolkit, we provide a comprehensive evaluation protocol for evaluating the robustness of defense methods against various attacks, and a novel metric called equilibrium defense success rate ($EDSR$). The protocol includes explaining the usage of MAD datasets and the backbone network training scheme. Primarily, we introduce the Meta-AT algorithm, which demonstrates high robustness against learned, unknown, online, and offline adversarial attacks after few-shot learning. In this study, Meta-AT serves as the baseline algorithm for the MAD benchmark, compared with other SOTA algorithms such as traditional AT [13], Fast-AT [22], Free-AT [23], ATTA [24], and YOPO [25]. Experimental results confirm the effectiveness and efficiency of Meta-AT. It requires approximately 0.33 min (MAD-M), 1 min (MAD-C), and 2 min (MAD-T) to achieve the highest values of SA and $EDSR$. At these points, the average $EDSR$ values across various attacks reach 98.23% (MAD-M), 100.92% (MAD-C), and 55.26% (MAD-T), respectively. It is worth noting that the MAD benchmark places a greater emphasis on comprehensive defense against various types of attacks by integrating both online and offline AT, learning from examples across

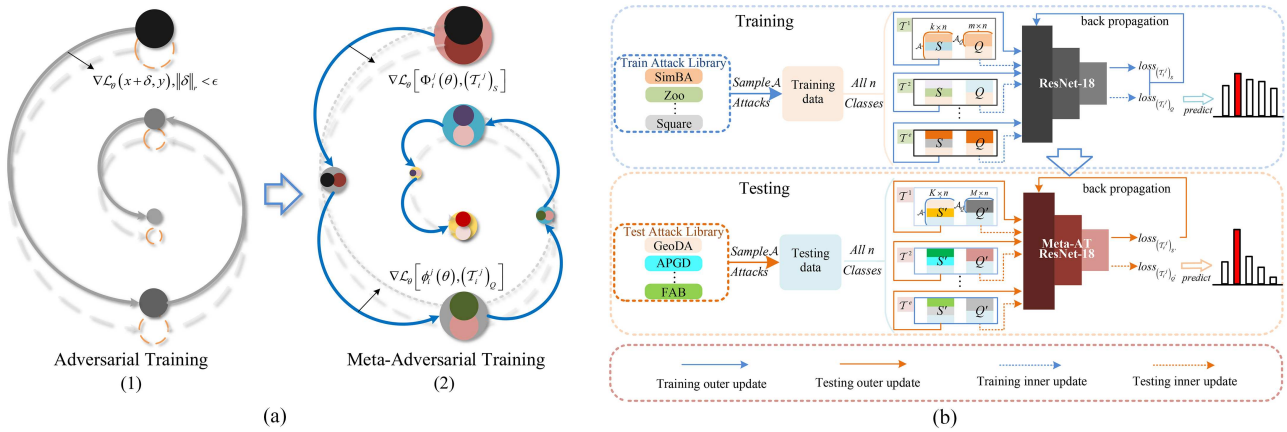


Figure 2 (Color online) (a) The diagram illustrates the updating of the model parameters θ under different training methods. The directional arrows indicate the convergence direction of the parameters, while the equations represent the associated gradients. In (1), the circles represent the current task parameters. In (2), the small circles with different colors within each task represent adversarial and clean examples, while the larger circles surrounding them represent the model parameters of the $\mathcal{T}$. For comparison, the long dashed line represents conventional training, and the short dashed line represents AT. (b) The details of the "$\mathcal{A}$-*way*, K-*shot*" Meta-AT. Different attacks in the attack library are represented by different colors, and the AEs in S, Q, S' and Q' have the same color as their corresponding attacks. ResNet-18 is used as an example for the target network.

different models, and significantly mitigating the "blind spots" present in existing research.

As depicted in Figure 2(a), we illustrate the basic principles of model parameter update methods used in AT and Meta-AT. In Meta-AT, we transform the single-step updates into internal and external updates. The AT is decomposed into multiple mini-ATs ($\mathcal{T}$s) as base tasks, enabling the model to learn from diverse AEs of fixed attack types in each epoch. Furthermore, the detailed framework for Meta-AT is depicted in Figure 2(b). In summary, our specific contributions can be summarized as follows.

- We present the first comprehensive MAD benchmark, which includes three extensive MAD datasets constructed by 30 mainstream adversarial attacks on MNIST, CIFAR-10, and Tiny-ImageNet, and an MAD evaluation toolkit.
- Our MAD evaluation protocol emphasizes offline AT to evaluate defense capabilities against unknown attacks through few-shot learning on MAD datasets. In particular, a more appropriate $EDSR$ metric is introduced to provide a comprehensive evaluation of defense robustness, post-defense clean sample accuracy, and learning capability.
- We propose a Meta-AT baseline algorithm feature with high robustness against learned, unknown, online, and offline adversarial attacks after few-shot learning. Compared with the SOTA adversarial defense methods, it achieves a high generalization defense capability while maintaining excellent SA. Extensive ablation studies of Meta-AT have also been demonstrated thoroughly.

The remainder of this article is organized as follows: In Section 2, related work is introduced. In Section 3, we state the rationale for our proposed MAD benchmark and the baseline algorithm Meta-AT. Experiments and results are presented in Section 4. The conclusion is provided in Section 5.

2 Related work

This section provides an overview of the fundamental concepts underpinning this paper, including AT-based adversarial defense methods and meta-learning-related approaches for adversarial defense.

2.1 Adversarial training

Adversarial defense methods based on AT are often viewed as a fundamental approach to defend against adversarial attacks, training models on both clean and adversarially generated examples to enhance robustness. Madry et al. [13] conceptualized AT into a non-convex and non-concave saddle-point problem, fitting it to the min-max problem as shown in (1). The inner maximization corresponds to an arbitrary adversarial attack.

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\delta\|_r<\epsilon} \mathcal{L}_\theta(x+\delta, y) \right], \tag{1}$$

where $\mathcal{L}_\theta(x+\delta, y)$ is the loss function of the network with parameter θ. (x, y) is the training pair sampled from the training set $\mathcal{D}$. δ is an adversarial perturbation constrained by its r-norm, which does not exceed the parameter ϵ.

Traditional AT suffers from certain drawbacks, including high computational cost, low generalization performance against unknown attacks, and degradation of the classification performance of the original model. To alleviate the computational burden, ATTA and YOPO are proposed which reduce the number of gradient calculations from outside and inside the network, respectively. These methods are influenced by the Free-AT and Fast-AT. To preserve the original classification ability, Farnia et al. [26] introduced an AT regularization technique utilizing spectral normalization. Many efforts are also dived into the generalization of AT, such as FLC pooling [27]. Unlike these approaches, which focus on enhancing specific aspects of AT or model robustness, our approach emphasizes continuous defense against a variety of attacks, aiming to achieve near-perfect comprehensive defense capabilities.

2.2 Meta-learning related adversarial defense methods

Meta-learning is particularly valuable in few-shot learning scenarios, as it aims to train models to quickly adapt to new tasks using limited data. It particularly addresses the *"$\mathcal{A}$-way, K-shot"* problem. A prominent method in this field is model-agnostic meta-learning (MAML) [28].

The integration of meta-learning with AEs is a burgeoning field. Goldblum et al. [29] developed the adversarial query (AQ) algorithm by replacing the query set in few-shot classification tasks with AEs. Liu et al. [30] proposed a long-term cross-adversarial training (LCAT) method that achieves similar effects to AQ but is more efficient. Qi et al. [31] introduced a meta-learning-based AT framework (MBATF) incorporating attention mechanisms to address cross-domain challenges in few-shot learning. These approaches focus on using adversarial defense strategies to improve the resilience and robustness of few-shot learning models.

Researchers have proposed innovative frameworks that leverage meta-learning for enhanced defense against adversarial attacks. Ma et al. [32] introduced a dual-network model for robust attack detection using limited examples, significantly advancing the field of the detection as defense. Metzen et al. [33] applied meta-learning to AT with IFGSM-generated meta-patches, effectively targeting universal patch attack defense. This approach represents a notable contribution towards defending against universal patch attacks in a targeted manner.

It is important to emphasize that we only focus on AT within the context of classification tasks, rather than on few-shot classification tasks or on enhancing robustness specifically for few-shot classification. The primary objective is to develop a classification model with high SA, capable of defending against known attacks while quickly adapting to defend against previously unknown attacks.

3 MAD benchmark

This section presents a comprehensive introduction to the MAD benchmark. It mainly focuses on three key aspects: the construction and characteristics of the MAD dataset, the MAD evaluation toolkit, and the Meta-AT baseline algorithm. Together, these components establish a standardized and versatile framework for meta-adversarial defense.

3.1 MAD dataset

The MAD dataset comprises three distinct datasets. Specifically, MAD-M and MAD-C are generated using a ResNet-18 model [34] trained on clean examples, which is subjected to 30 adversarial attacks on the MNIST and CIFAR-10 datasets. Similarly, MAD-T is generated using an EfficientNet-b0 model [35], also trained on clean examples, which undergoes 30 adversarial attacks on the Tiny-ImageNet dataset. The details of the 30 attacks, along with their corresponding IDs, can be found in Appendix A. Attacking both the training and testing examples in original datasets is time-consuming and not significant for the few-shot learning pattern. Therefore, we use the validation datasets from the original datasets for the attacks and obtain a large set of attacked datasets.

At the beginning of the dataset creation, we select attacks from two well-known and simple attack libraries, the Adversarial Robustness Toolbox (ART) [36] and the AdverTorch [37]. The SA values of the model in the original testing dataset are MNIST: 98.84%, CIFAR-10: 94.82%, and Tiny-ImageNet: 71.90%. To maintain the category balance of labels, we include all classes in datasets for training and testing (10 classes for MAD in each episode). The RA values of the original validation datasets after various attacks are given by the histogram in Appendix B. As can be seen, not all attacks are effective and their effectiveness varies across datasets. Generally, larger image sizes in the dataset result in more pronounced attack effects. Some attacks do not have sufficient examples, so specific attacks are removed from the experiments. Thus, attacks 0, 1, 2, 9, 10, 12, 24, and 28 from MAD-M, attack 0 from MAD-C, and attacks 12 and 29 from MAD-T are removed. For instance, class 8 in the 24th attack stands

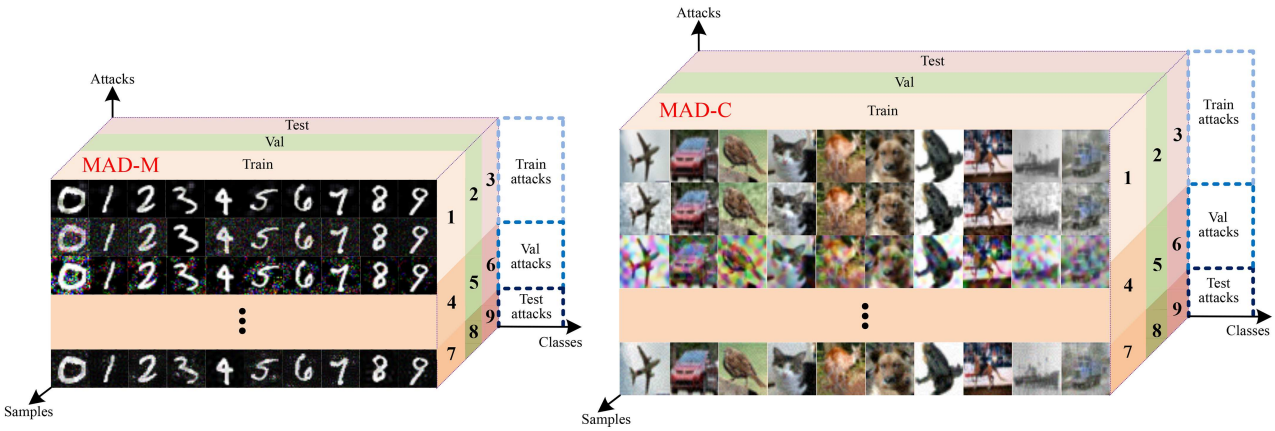


Figure 3 (Color online) The details and simple examples of MAD-M and MAD-C.

out as it has a very low number of AEs and results in an uneven distribution of examples among the categories. Due to this imbalance, we have to abandon these attacks.

Careful screening is applied to retain only successfully attacked examples under each attacked dataset. This results in each attacked dataset having an RA of 0% on the initially trained clean model. MAD-M consists of 10 classes with a single image size of 28×28 under 22 attacks, MAD-C consists of 10 classes with a single image size of 32×32 under 29 attacks, while MAD-T consists of 200 classes with a single image size of 64×64 under 28 attacks. It can be seen that the datasets we used are more comprehensive. The training, validation, and testing sets are partitioned according to the partitioning pattern of few-shot learning, taking into account the year and perturbation measurements proposed by different attacks. For better evaluation, examples from the original testing dataset and the MAD datasets are split in a $3:1:1$ ratio for each class. Figure 3 shows the two types of MAD datasets obtained, with legends added to improve readability and usability.

3.2 MAD evaluation toolkit

3.2.1 *Evaluation protocol*

Traditional AT evaluation protocols primarily assess models based on black-box/white-box attacks, perturbation size, and attack diversity. Our MAD benchmark evaluation protocol significantly extends these aspects. It offers a broader range of attack categories under both black-box and white-box settings, along with detailed perturbation sizes and hyperparameter configurations for different attacks (see Appendix A for details). In addition, we introduce a novel training and testing framework for meta-adversarial defense, providing comprehensive guidelines for evaluating defense methods on newly curated datasets specifically designed for Meta-AT. We aim to foster a deeper understanding of model resilience against a diverse set of adversarial threats, ensuring that models are not only robust but also generalize well across different attack scenarios.

To facilitate the subsequent utilization of the dataset and the reproducibility of the MAD benchmark, we color-coded and numerically labeled the side profile of the dataset model, as shown in Figure 3. Specifically, labels "1" and "2" correspond to samples under "Train attacks" used during Meta-AT training; labels "4" and "5" represent samples under "Val attacks" used for Meta-AT validating; labels "7" and "8" are the samples under the "Test attacks" used for testing the rapid fine-tuning capabilities of models pre-trained with Meta-AT; labels "3" and "6" denote unknown examples designed to test the defensive capabilities of the Meta-AT-trained model against "Learned attacks"; label "9" includes unknown examples intended to assess the defensive capabilities of Meta-AT-trained models against entirely "Unknown attacks". The SA values of all the robust models are implemented on clean examples corresponding to "9".

3.2.2 *Evaluation metrics*

To showcase the effectiveness of the proposed Meta-AT approach, the performance can be assessed and compared using the following metrics.

- ***SA*** evaluates the accuracy of correctly classifying clean examples across various scenarios, serving as a baseline metric for assessing general classification performance.
- ***RA*** measures the classification accuracy on adversarial examples generated under various attacks, reflecting robustness against adversarial perturbations.

• **AutoAttack [38] (*AA*)** is a widely recognized benchmark for assessing robustness against adversarial attacks. Its standardized nature makes it a reliable metric for comparing different defense strategies.

• **Operating time (*OT*)** quantifies the time required for the robust AT approaches to reach the highest RA, reflecting computational efficiency.

• ***EDSR*** assesses the equilibrium between defense capability and performance on clean sample classification after rapid learning. A high $EDSR$ indicates effective robustness with minimal degradation in primary classification accuracy, showcasing strong few-shot learning capabilities. Moreover, $EDSR$ serves as an early warning indicator for potential challenges, such as catastrophic overfitting [22] and robust overfitting [39], which might arise when learning from limited examples. The specific formula for $EDSR$ is presented as.

$$EDSR = \frac{RA^* - RA}{SA - RA + \alpha\Delta}, \tag{2}$$

where RA and RA^* denote the robust accuracies of the model before and after adversarial defense, respectively. The term $\Delta = (SA - SA^*)$ quantifies the performance gap between the original baseline SA and the SA achieved after few-shot learning (SA^*). The parameter $\alpha = 1$ is a tunable weight that highlights the impact of clean sample performance degradation resulting from adversarial defense training. When Meta-AT is not applied, the penalty term $\Delta = 0$, and Eq. (2) reduces to the traditional definition of the defense success rate (DSR). Therefore, the proposed metric is versatile and can be applied to evaluate various defense methods.

3.3 Meta-AT baseline algorithm

Why Meta-AT works? The decomposition of the optimization problem into multi-step learning can help prevent overfitting and facilitate synthesis learning. Progressive networks [40] exemplify this by simplifying subtasks before tackling complex ones, thereby avoiding catastrophic forgetting and enhancing learning efficiency. The PGD attack, an iterative improvement over the fast gradient sign method (FGSM) [41] attack, aims for stronger adversarial effects by maximizing model loss. Meta-learning decomposes tasks into smaller subtasks, improving adaptability and transferability across domains. Therefore, Meta-AT can be recursively rolled out to split a single AT task into multiple tasks, enhancing robustness against various attacks while retaining primary task accuracy. The basic configuration of Meta-AT typically involves the following elements.

• **Task $\mathcal{T}$.** MAD datasets with a specific data distribution $\mathcal{P}$ are partitioned into distinct tasks $\mathcal{T}$ based on the categorization of attacks for few-shot learning.

• **Support (S) and Query (Q) set.** Each $\mathcal{T}$ comprises an attacked S set and an attacked Q set. The S set is employed for fine-tuning the model parameters during the inner loop, while the Q set is utilized to evaluate the model's performance during the outer loop.

• **Way.** In the context of Meta-AT, "way" refers to the number of fixed attack categories $\mathcal{A}$ associated with each $\mathcal{T}$ for learning.

• **Shot.** It denotes the number of AEs sampled for each "way" in the Q set.

• **Episode (e).** The variable e denotes the number of tasks $\mathcal{T}$ per epoch in the entire learning process.

The comprehensive details of Meta-AT are illustrated in Figure 2. In the training phase, the meta-task of each batch is a $\mathcal{T}$. S set comprises AEs generated using a set of $\mathcal{A}$ train attacks, while Q set includes AEs generated by $\mathcal{A}_Q$ attacks, which are randomly selected from $\mathcal{A}$ train attacks. Data balancing is performed by selecting a specific number of AEs under all n classes per attack ($k \times n$) for S set, and a specific number of AEs per attack for Q set ($m \times n$), ensuring no intersection between S and Q sets. The RA^* evaluated on the query set (Q) represents the model's comprehensive robustness after adapting to multiple tasks $\mathcal{T}$.

During the validating/testing phase, testing sets S' and Q' are also not overlapped, and they are selected from ($K \times n \times \mathcal{A}$) examples and ($M \times n \times \mathcal{A}_Q$) examples under all n classes, respectively. Importantly, $\mathcal{A}_Q$ validating/testing attacks are distinct from $\mathcal{A}$ train attacks, signifying a key setup in which the Meta-AT algorithm aims to learn from known attacks to adapt models to unknown attacks. More specifically, S'/S set is used for fine-tuning and Q'/Q set for evaluating. This configuration parallels the few-shot learning classification task, referred to as *"$\mathcal{A}$-way, K-shot"*. Additional parameter settings can be found in Table 1, and the optimized target of Meta-AT is shown as

$$\min_{\theta} \mathbb{E}_{\mathcal{T}\sim\mathcal{P}(\mathcal{T})}[\mathcal{F}(\theta)], \tag{3}$$

where $\mathcal{F}(\theta) = \mathcal{L}(\theta) + \eta\mathcal{R}(\theta)$, and $\mathcal{L}(\theta)$ is the classification loss function defined as

$$\mathcal{L}(\theta) = \frac{1}{e}\sum_{j=1}^{e} \max_{\|\delta_b\|_r < \epsilon} \ell_c\left(f_\theta(x_j + \delta_b), y_j\right), \tag{4}$$

Table 1 Parameters of Meta-AT.

Parameter	Default	Description
β	0.01	Inner update learning rate.
λ	0.001	Outer update learning rate.
η	1	Coefficient of SAR.
E	10	The maximum number of training epochs.
e	100	The maximum number of training episodes in each epoch.
B	1	The batch size in each episode.
p	20	Patience index, the number of the episodes in early stopping.
$\mathcal{A}$	5	Shot-attack-way, the number of attacks used in S/S' set.
$\mathcal{A}_Q$	1	Query-attack-way, the number of attacks used in Q/Q' set.
k	15	Train-shot, the number of examples sampled per class in S set during training.
m	6	Train-query, the number of examples sampled per class in Q set during training.
K	1	Val/test-shot, the number of examples sampled per class in S' set during validating/testing.
M	15	Val/test-query, the number of examples sampled per class in Q' set during validating/testing.

where $\ell_c(\cdot,\cdot)$ is the cross-entropy loss function and $\eta > 0$ is a tuning parameter. $\mathcal{R}(\theta)$ is the smoothness-inducing adversarial regularizer (SAR) applied during the fine-tuning phase to effectively control the complexity of the model. We define $\mathcal{R}(\theta)$ as follows:

$$\mathcal{R}(\theta) = \frac{1}{e}\sum_{j=1}^{e} \max_{\|\delta_b\|_r<\epsilon} \ell_k\left(f_\theta(x_j+\delta_b), f_\theta(x_j)\right), \tag{5}$$

where ℓ_k is chosen as the symmetrized KL-divergence. Specifically, given the model $f(\cdot;\theta)$, the datapoints for the e target $\mathcal{T}$s in each epoch are represented as $\left\{(x_j,y_j)^d \mid j\in[1,e], d\in[1,n]\right\}$.

$$f(x;\theta) = \Phi_{\phi(\theta,mAT_a)}(x). \tag{6}$$

Note that $\Phi(\theta)$ is a model represented by a parameterized function. We expect to obtain the best θ at the time of minimum loss as the fine-tuning model $\phi(\theta,\mathcal{T}_a)$ under the ath attack adapts to the new $\mathcal{T}_b$ under the bth attack. The training phase of Meta-AT is detailed in Algorithm 1, while the testing phase is outlined in Algorithm 2.

Algorithm 1 Meta-AT (training phase).

Input: Pre-trained model ($\Phi(\theta)$), training dataset (T) in MAD, number of attacks in S ($\mathcal{A}$) and Q ($\mathcal{A}_Q$) sets, learning rates (β and λ), regularization parameter (η), total epochs (E).
Output: Fine-tuned model ($\phi(\theta)$).

Randomly initialize parameter θ and split T set into S and Q sets;
while *epoch* $\leqslant E$ **do**
  Sample batch of tasks $\{\mathcal{T}_i\}_{i=1}^{\mathcal{A}} \sim \mathcal{P}(\mathcal{T})$, where $\mathcal{T}_i = \left\{\left(\mathcal{T}_i^j\right)_S, \left(\mathcal{T}_{i'}^j\right)_Q \mid j\in[1,e],\ i'\in[1,\mathcal{A}_Q]\right\}$;
  for all $\mathcal{T}_i$ **do**
    Sample $k\times n$ datapoints $\mathcal{D}_S = \left\{\left(x_i^j, y_i^j\right)\right\}_{d=1}^{n}$ from $\left(\mathcal{T}_i^j\right)_S$;
    Evaluate $\nabla_\theta \mathcal{F}_{\left(\mathcal{T}_i^j\right)_S}(\theta)$ using $\mathcal{D}_S$ and $\mathcal{F}_{\left(\mathcal{T}_i^j\right)_S}$ in (4)–(6);
    Compute adapted parameters with gradient descent: $(\theta_i^j)' = \theta - \beta\nabla_\theta \mathcal{F}_{\left(\mathcal{T}_i^j\right)_S}(\theta)$;
    Sample $K\times n$ datapoints $\mathcal{D}_Q = \left\{\left(x_{i'}^j, {}^j_{i'}\right)\right\}_{d=1}^{n}$ from $(\mathcal{T}_{i'}^j)_Q$ for the meta-update;
  end for
  Update $\theta \leftarrow \theta - \frac{\lambda}{e}\sum_{\{\mathcal{T}_i\}_{i=1}^{\mathcal{A}}\sim\mathcal{P}(\mathcal{T})} \text{Clip}\left\{\nabla_\theta \mathcal{F}_{\left(\mathcal{T}_{i'}^j\right)_Q}\left[(\theta_i^j)'\right], 1\right\}$ using $\mathcal{D}_Q$ and $\mathcal{F}_{\left(\mathcal{T}_{i'}^j\right)_Q}$ in (4)–(6);
end while

4 Experiment

In this section, we present the basic configurations of the five SOTA defense methods. Subsequently, we validate that Meta-AT maintains high SA and robustness compared to these methods under traditional evaluation protocols. Furthermore, we compare these methods with Meta-AT, demonstrating the authority of Meta-AT in the field of meta-adversarial defense. Finally, we enhance our understanding by conducting ablation studies to examine the impact of key parameter configurations in the algorithm. All experiments are performed on an HPC server equipped with four Nvidia Tesla P100 12G GPUs and Intel (R) Xeon Gold 6132 CPU.

Algorithm 2 Meta-AT (testing phase).

Input: Fine-tuned model ($\phi(\theta)$), validating/testing dataset (T') in MAD, number of attacks in S' ($\mathcal{A}$) and Q' ($\mathcal{A}_Q$) set, learning rates (β' and λ') of saved best checkpoint.
Output: Fine-tuned model ($\phi(\theta)'$).
1: Split T' set into S' and Q' sets;
2: Sample batch of tasks $\{\mathcal{T}_i\}_{i=1}^{\mathcal{A}} \sim \mathcal{P}(\mathcal{T})$, where $\mathcal{T}_i = \left\{ \left(\mathcal{T}_i^j\right)_{S'}, \left(\mathcal{T}_{i'}^j\right)_{Q'} \mid j \in [1, e],\ i' \in [1, \mathcal{A}_Q] \right\}$;
3: **for** all $\mathcal{T}_i$ **do**
4: Sample $k \times n$ datapoints $\mathcal{D}_{S'} = \left\{ \left(x_i^j, y_i^j\right) \right\}_{d=1}^{n}$ from $\left(\mathcal{T}_i^j\right)_{S'}$;
5: Evaluate $\nabla_\theta \mathcal{F}_{\left(\mathcal{T}_i^j\right)_{S'}}(\theta)$ using $\mathcal{D}_{S'}$ and $\mathcal{F}_{\left(\mathcal{T}_i^j\right)_{S'}}$ in (4)–(6);
6: Compute adapted parameters with gradient descent: $(\theta_i^j)' = \theta - \beta' \nabla_\theta \mathcal{F}_{\left(\mathcal{T}_i^j\right)_{S'}}(\theta)$;
7: Sample $K \times n$ datapoints $\mathcal{D}_{Q'} = \left\{ \left(x_{i'}^{j,j}\right) \right\}_{d=1}^{n}$ from $(\mathcal{T}_{i'}^j)_{Q'}$ for the meta-update;
8: **end for**
9: Update $\theta \leftarrow \theta - \frac{\lambda'}{\epsilon} \sum_{\{\mathcal{T}_i\}_{i=1}^{\mathcal{A}} \sim \mathcal{P}(\mathcal{T})} \text{Clip} \left\{ \nabla_\theta \mathcal{F}_{\left(\mathcal{T}_{i'}^j\right)_{Q'}} \left[(\theta_i^j)' \right], 1 \right\}$ using $\mathcal{D}_{Q'}$ and $\mathcal{F}_{\left(\mathcal{T}_{i'}^j\right)_{Q'}}$ in (4)–(6) with early-stopping;

4.1 Experimental results compared with SOTA methods

4.1.1 *Basic settings*

We select five SOTA defense methods based on comprehensive reviews from three aforementioned comprehensive references [14–16] in Section 1. These methods employ diverse techniques and configurations.

• **Robust-AT (R-AT)** is a PGD-based AT defense method including the following robust training techniques. These include eval mode batch normalization (BNeval) and early stopping for generating AEs. The key hyperparameters vary by dataset: for CIFAR-10 and Tiny-ImageNet, training involves 110 epochs, a batch size of 128, a multistep learning rate schedule, and the SGD optimizer with an initial learning rate of 0.1 and a weight decay of 5×10^{-4}; for MNIST, training is conducted over 15 epochs with a batch size of 64 and a initial learning rate of 0.01. All the methods are under the ℓ_∞ threat model of maximal perturbation $\epsilon = \frac{8}{255}$, without accessibility to additional data. Retaining the original configuration, these robust training techniques are consistently applied across all the following methods, which is why an "R" is prefixed to their original names. Consequently, the following descriptions focus exclusively on the specific parameters distinguishing each method. Notably, R-AT also incorporates moderate label smoothing (LS) following the TRADES framework.

• **Robust-Fast-AT (R-Fast)** trains for 10 epochs on MNIST and 15 epochs on CIFAR-10, with a batch size of 128. It employs a cyclic learning rate schedule, setting the minimum learning rate ($lr_{\min}$) to 0 and the maximum learning rate ($lr_{\max}$) to 0.005 for MNIST and $lr_{\max} = 0.2$ for CIFAR-10.

• **Robust-Free-AT (R-Free)** trains over 10 epochs with a batch size of 128. A cyclic learning rate schedule is also used, where $lr_{\min} = 0$ and $lr_{\max} = 0.04$ is set to CIFAR-10 and MNIST. Additionally, it includes minibatch replays of 8.

• **Robust-YOPO (R-YOPO)** integrates the TRADES-YOPO method, using TRADES-YOPO-5-10 for 40 epochs on the MNIST and TRADES-YOPO-5-3 for 38 epochs on CIFAR-10[1)].

• **Robust-ATTA (R-ATTA)** adopts the TRADES framework, performing 40 attack iterations (TRA40) over 60 epochs on MNIST and 10 attack iterations (TRA10) over 38 epochs on CIFAR-10[1)].

• **Meta-AT (ours)** adopts an episodic training framework with 100 episodes per epoch. The training process typically spans 10 epochs for MAD-C and MAD-T, and 5 epochs for MAD-M. Further detailed parameter settings are provided in Table 1, and additional training techniques are discussed in the "Training tricks" in Subsection 4.2.

The RA^* values of the SOTA models on the MAD-C are illustrated in Figure 1. The results indicate that all robust models are affected by offline AEs generated from the original model, with varying degrees of impact depending on the attack and the model. Apart from the traditional R-AT, other robust models are more susceptible to offline attacks. Notably, R-Fast, R-ATTA, and R-YOPO exhibit particularly poor performance in this regard.

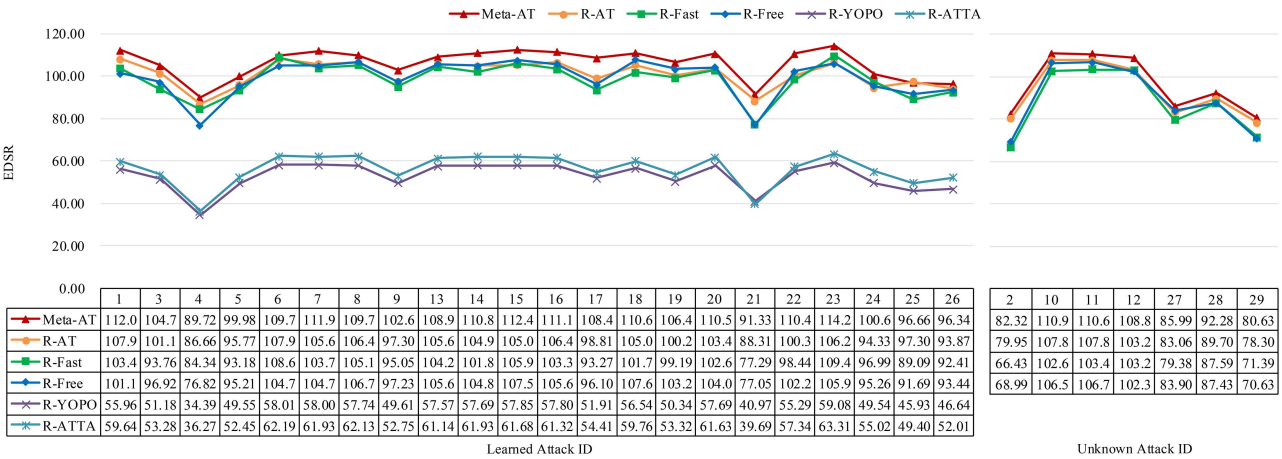
4.1.2 *Clean-sample classification accuracy and robustness*

The two primary criteria in traditional evaluation protocols to evaluate the robustness of a model against AEs are its SA and the accuracy under standard strong attack AA. For EfficientNet, we implemented it only on the traditional R-AT and our Meta-AT. The basic tricks used on ResNet also do not transfer well to the larger EfficientNet models and Tiny-ImageNet. Nevertheless, our focus is on proposing a continual learning defense framework based

1) The abbreviations in quotation marks are the optimal configuration names in the original manuscripts; see the original article for more basic hyperparameter settings.

Table 2 The evaluation results (%) of different adversarial defense algorithms on the clean dataset. The best results are in bold.

Method	MAD-M		MAD-C		MAD-T	
	SA	AA	SA	AA	SA	AA
No-defense	98.84	0.00	94.82	0.00	71.90	0.00
R-AT [13]	99.40	**99.10**	88.35	**52.20**	48.65	**10.28**
R-Fast [17]	98.10	97.60	79.40	41.40	–	–
R-Free [16]	99.10	98.70	80.40	42.80	–	–
R-YOPO [15]	97.93	93.31	80.25	40.62	–	–
R-ATTA [14]	98.83	94.19	79.90	40.90	–	–
Meta-AT (ours)	**99.65**	98.88	**89.00**	46.00	**52.80**	9.86



	1	3	4	5	6	7	8	9	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Meta-AT	112.0	104.7	89.72	99.98	109.7	111.9	109.7	102.6	108.9	110.8	112.4	111.1	108.4	110.6	106.4	110.5	91.33	110.4	114.2	100.6	96.66	96.34
R-AT	107.9	101.1	86.66	95.77	107.9	105.6	106.4	97.30	105.6	104.9	105.0	106.4	98.81	105.0	100.2	103.4	88.31	100.3	106.2	94.33	97.30	93.87
R-Fast	103.4	93.76	84.34	93.18	108.6	103.7	105.1	95.05	104.2	101.8	105.9	103.3	93.27	101.7	99.19	102.6	77.29	98.44	109.4	96.99	89.09	92.41
R-Free	101.1	96.92	76.82	95.21	104.7	104.7	106.7	97.23	105.6	104.8	107.5	105.6	96.10	107.6	103.2	104.0	77.05	102.2	105.9	95.26	91.69	93.44
R-YOPO	55.96	51.18	34.39	49.55	58.01	58.00	57.74	49.61	57.57	57.69	57.85	57.80	51.91	56.54	50.34	57.69	40.97	55.29	59.08	49.54	45.93	46.64
R-ATTA	59.64	53.28	36.27	52.45	62.19	61.93	62.13	52.75	61.14	61.93	61.68	61.32	54.41	59.76	53.32	61.63	39.69	57.34	63.31	55.02	49.40	52.01

	2	10	11	12	27	28	29
Meta-AT	82.32	110.9	110.6	108.8	85.99	92.28	80.63
R-AT	79.95	107.8	107.8	103.2	83.06	89.70	78.30
R-Fast	66.43	102.6	103.4	103.2	79.38	87.59	71.39
R-Free	68.99	106.5	106.7	102.3	83.90	87.43	70.63

Figure 4 (Color online) The performance (%) of different adversarial defense algorithms under different attacks on MAD-C.

on existing robust models. Therefore, despite the lower AA of EfficientNet after AT, it remains valid for the experimental verification.

Table 2 presents SA and AA values of seven models on three MAD datasets. It can be seen that most AT methods trade off some accuracy for higher robustness. However, Meta-AT achieves higher accuracy than the original SA on MAD-M, mainly because the learning mode of Meta-AT turns offline AEs into fine-grained data augmentation. Overall, the key advantage of Meta-AT is its ability to achieve both high SA and high model robustness.

4.1.3 *Defense capability and continual learning*

Figure 4 presents the $EDSR$ results for various defense methods under different attacks. Our study primarily examines the defense capabilities and generalization to unknown attacks of the original model by modifying the training framework without changing the model structure. For the comprehensive evaluation metric $EDSR$, when $\Delta = 0$, $EDSR$ simplifies to the traditional DSR. Therefore, for R-YOPO and R-ATTA, we only provide DSR under different attacks.

In the testing mode of Meta-AT, $EDSR$ further evaluates the learning ability of the model after training on a small number of AEs. When the learned Δ is larger, $EDSR$ increases, and if the post-learning SA^* exceeds the pre-learning SA, $EDSR$ can exceed 100%. As shown in Figure 4, models trained with Meta-AT often outperform other methods, with 20 values exceeding 1. The counts for R-AT, R-Free, and R-Fast are 14, 16, and 14, respectively, indicating similar learning abilities. Combining this figure with Table 2, the offline and online combination mode of Meta-AT enhances the defense effectiveness of various methods. Meta-AT outperforms other methods on learned attacks except for attack 25 and demonstrates higher generalization to unknown attacks.

Table 3 presents the average values of SA^*, AA, and $EDSR$ of different defense methods on MAD-C and MAD-M. Meta-AT maintains the highest SA^*, with some improvements, while AA slightly decreases. In terms of $EDSR$, Meta-AT scores the highest, followed by R-AT, R-Fast, and R-Free. In summary, models trained with Meta-AT are effective in defending against both learned and unknown attacks, demonstrating high robustness and continual learning capability.

Table 3 The evaluation results (%) of different adversarial defense algorithms on MAD-C. The best results are in bold.

Defense method		R-AT	R-Fast	R-Free	Meta-AT (ours)
MAD-M	SA^*	99.00	98.45	98.90	**99.76**
	AA	**98.89**	97.46	98.62	98.83
	$EDSR$	88.61	89.80	86.39	**98.23**
MAD-C	SA^*	89.00	79.45	76.40	**91.76**
	AA	**47.19**	39.46	40.62	45.83
	$EDSR$	96.85	93.03	94.39	**100.92**
MAD-T	SA^*	49.00	–	–	**53.76**
	AA	**9.89**	–	–	9.83
	$EDSR$	48.61	–	–	**55.26**

Table 4 The evaluation results (%) of different learning rate configurations on MAD-C. The best results are in bold.

β	λ	SA^*	$EDSR$
0.1	0.01	NaN	NaN
0.01	0.001	**95.50**	**100.92**
0.001	0.0001	93.20	97.80

4.2 Ablation study

To maximize the advantages of the proposed Meta-AT, we performed comprehensive ablation studies on several critical aspects, including the classification backbone network, the *"$\mathcal{A}$-way, K-shot"* setting, and the patience index p. These experiments aimed to evaluate the flexibility, adaptability, and overall performance of Meta-AT under various configurations. The results provide actionable insights into the robustness and efficiency of the Meta-AT approach across different scenarios and attack types.

4.2.1 *Meta-step learning rate*

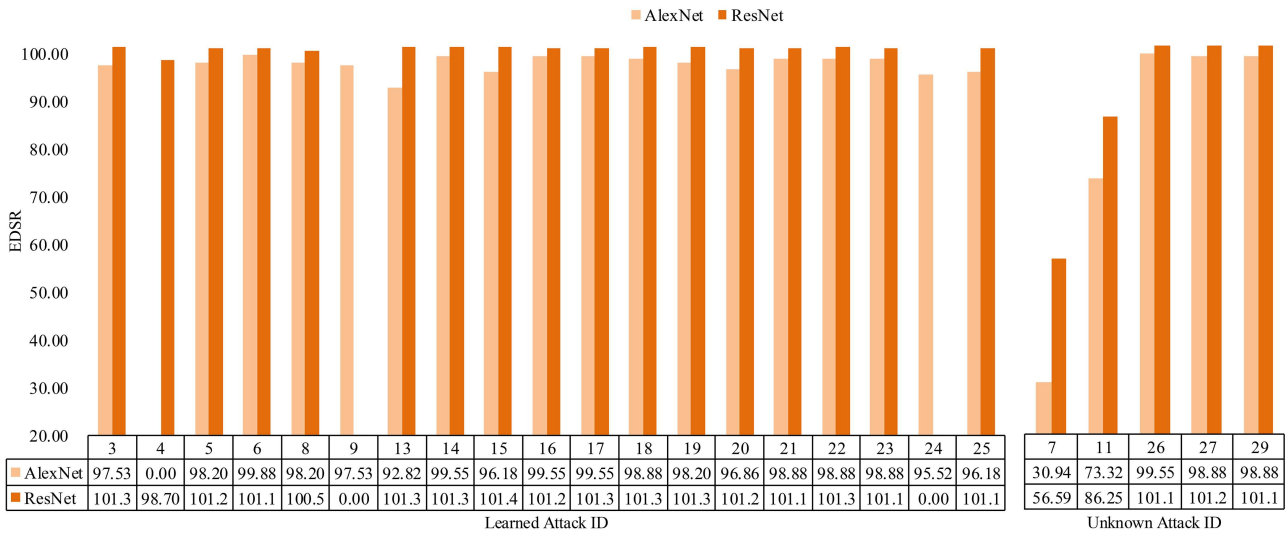
The selection of learning rates has a significant impact on both the convergence behavior of the network and its final performance. Inspired by meta-learning, we opted against using excessively large learning rates to ensure stability during training. Table 4 summarizes the effects of different learning rate configurations on MAD-C performance. As shown in the table, the combination of $\beta = 0.01$, $\lambda = 0.001$ yields the most balanced performance, achieving competitive results for both SA^* and $EDSR$.

4.2.2 *Classification backbone network*

To assess the effectiveness of our proposed Meta-AT on different classification backbone networks, we also created the MAD-M dataset based on AlexNet [42]. Using the aforementioned parameters, Figure 5 shows the $EDSR$s of different attacks after applying AEs fine-tuned on both AlexNet and ResNet-18 under the Meta-AT in a "5-*way*, 1-*shot*" setting. The graph indicates that attack 4 is ineffective on AlexNet, whereas attacks 9 and 24 are ineffective on ResNet-18. This discrepancy can be attributed to the relatively simple distribution of the MNIST dataset, which makes the simpler AlexNet network better suited for MNIST and most attacks effective against it while encountering fewer ineffective attacks. In the face of unknown attacks, the ResNet-18-based $EDSR$s consistently outperform those of AlexNet. This is attributed to the higher complexity and larger capacity of ResNet-18, resulting in higher SA^*. For unknown attacks, attacks 7 and 11 exhibit strong adversarial characteristics, and ResNet-18 demonstrates superior adversarial defense capabilities compared to AlexNet. Overall, these experiments demonstrate the transferability of adversarial defense via Meta-AT across different classification backbone networks. In the previous comparison experiments, we standardized the backbone network to the higher capacity ResNet-18.

4.2.3 *Training hyperparameters*

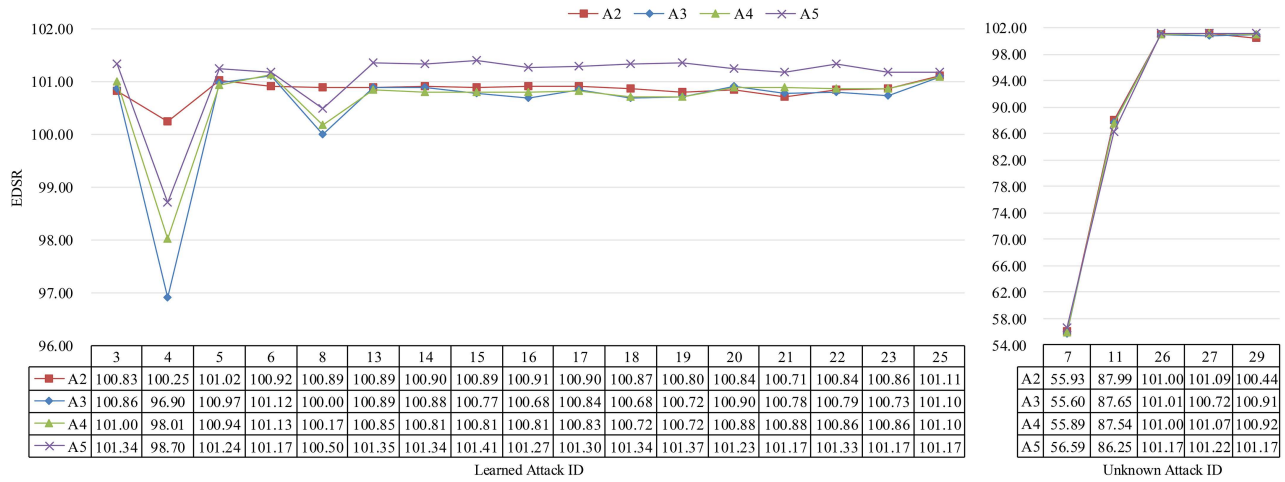
"$\mathcal{A}$-way, K-shot". This paper mainly studies a learnable model under the *"$\mathcal{A}$-way, K-shot"* mode by using AEs of $\mathcal{A}$ known attacks. New attacks can achieve a higher defense rate by fine-tuning this model with K examples. Among them, the value of K is derived from the context of few-shot classification tasks. Hence, we need to study the setting of parameter $\mathcal{A}$. The selection of $\mathcal{A}$ as 2, 3, 4, and 5 is based on MAD-M, excluding $\mathcal{A} = 1$ due to its similarity to the traditional AT. Figure 6 displays the $EDSR$s for different Meta-AT models ($\mathcal{A} = 2, 3, 4, 5$ correspond to A2, A3, A4, and A5 in the figure). It can be observed that the impact of $\mathcal{A}$ on model training is



	3	4	5	6	8	9	13	14	15	16	17	18	19	20	21	22	23	24	25
AlexNet	97.53	0.00	98.20	99.88	98.20	97.53	92.82	99.55	96.18	99.55	99.55	98.88	98.20	96.86	98.88	98.88	98.88	95.52	96.18
ResNet	101.3	98.70	101.2	101.1	100.5	0.00	101.3	101.3	101.4	101.2	101.3	101.3	101.3	101.2	101.1	101.3	101.1	0.00	101.1

	7	11	26	27	29
AlexNet	30.94	73.32	99.55	98.88	98.88
ResNet	56.59	86.25	101.1	101.2	101.1

Figure 5 (Color online) The performance (%) of Meta-AT under different backbone networks on MAD-M.



	3	4	5	6	8	13	14	15	16	17	18	19	20	21	22	23	25
A2	100.83	100.25	101.02	100.92	100.89	100.89	100.90	100.89	100.91	100.90	100.87	100.80	100.84	100.71	100.84	100.86	101.11
A3	100.86	96.90	100.97	101.12	100.00	100.89	100.88	100.77	100.68	100.84	100.68	100.72	100.90	100.78	100.79	100.73	101.10
A4	101.00	98.01	100.94	101.13	100.17	100.85	100.81	100.81	100.81	100.83	100.72	100.72	100.88	100.88	100.86	100.86	101.10
A5	101.34	98.70	101.24	101.17	100.50	101.35	101.34	101.41	101.27	101.30	101.34	101.37	101.23	101.17	101.33	101.17	101.17

	7	11	26	27	29
A2	55.93	87.99	101.00	101.09	100.44
A3	55.60	87.65	101.01	100.72	100.91
A4	55.89	87.54	101.00	101.07	100.92
A5	56.59	86.25	101.17	101.22	101.17

Figure 6 (Color online) The performance (%) of Meta-AT under different ways on MAD-M.

limited, with notable fluctuations occurring in only a few Meta-ATs for known attacks (e.g., attacks 4 and 8) and less impact on unknown attacks. Figure 6 alone is insufficient to determine a suitable $\mathcal{A}$ value. Therefore, we comprehensively consider the average values of SA^* and $EDSR$ across different models to differentiate them. They are A2 (99.50 & 95.07), A3 (99.55 & 94.87), A4 (**99.65** & 94.98), and **A5 (99.65 & 95.18)**. When $\mathcal{A} = 5$, the highest SA^* and $EDSR$ are observed (shown in bold). In conclusion, the *"5-way, 1-shot"* pattern serves as the basis for comparative experiments.

Patience index (p). The OT of Meta-AT is primarily impacted by the number of episodes, as evident from the 100 episodes testing process of Meta-AT for a single attack shown in Figure 7. The graph reveals significant fluctuations during the initial learning stage. The overall trend of learning is characterized by an initial decrease, followed by an increase, and then another decrease. The initial decrease to increase reflects the pattern of AT, while the subsequent decrease indicates the gradual overfitting of the model. To prevent model overfitting and improve $EDSR$, an appropriate value of p in the early stopping mechanism is crucial. Learning over 30 steps for different attacks tends to stabilize, taking about 1 s per step. Based on experimental experience, we set $p = 20$ to ensure the effectiveness of Meta-AT. The OT of Meta-AT, learning a few AEs requires approximately 0.33 min for MAD-M, 1 min for MAD-C, and 2 min for MAD-T allowing the model to achieve the highest SA^*. Overall, Meta-AT significantly shortens the learning period, and combined with its high $EDSR$ and SA^* establishes itself as a leading method in industrial-grade online AT-based defense methods.

Training tricks. In addition to the core Meta-AT framework, we explored various training tricks to further enhance the performance and robustness of the model. These tricks include integrating strategies such as data

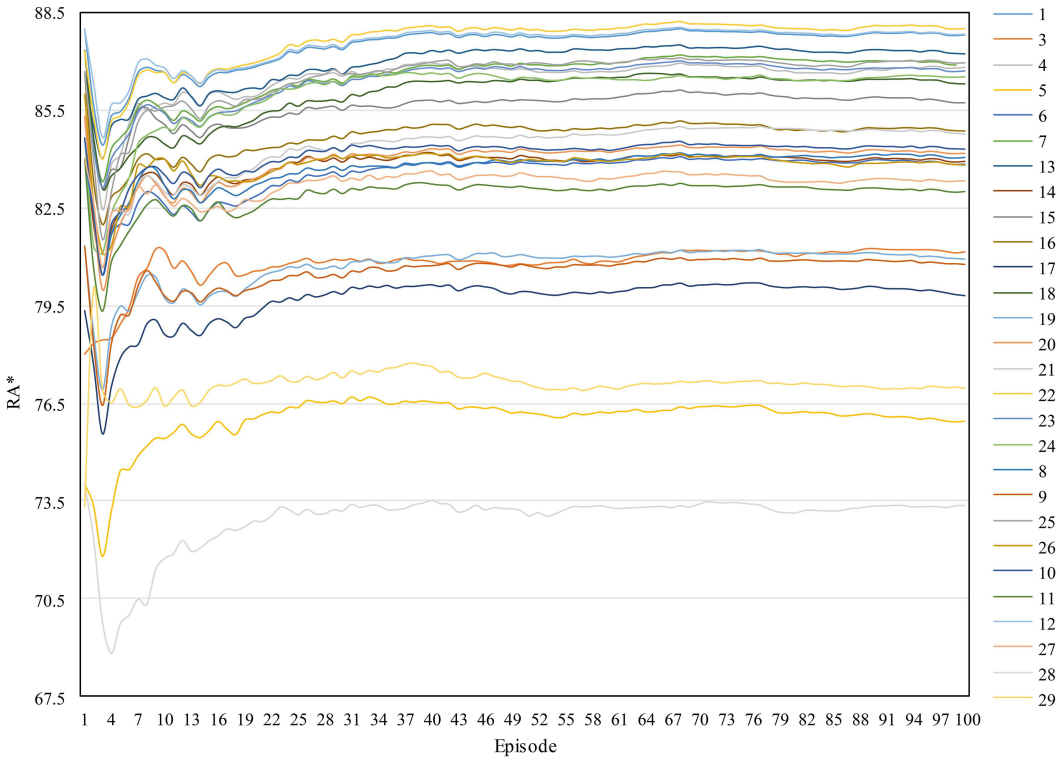


Figure 7 (Color online) Robust classification accuracy following adversarial defense (RA^*, %) during 5-*way*, 1-*shot* Meta-AT testing on MAD-C test examples.

Table 5 The evaluation results (%) under different training tricks of Meta-AT on MAD-C. The best results are in bold.

Tricks					Metrics			Diff
SAR	GC	WLR	LS	Mixup	$EDSR$	SA^*	AA	
N	N	N	N	N	97.70	**98.30**	32.60	–
Y	N	N	N	N	98.30	97.70	36.50	3.90 ↑
N	Y	N	N	N	99.79	96.60	37.70	5.10 ↑
N	N	Y	N	N	98.66	97.60	35.60	3.00 ↑
N	N	N	Y	N	98.89	97.20	32.70	0.10 ↑
N	N	N	N	Y	99.26	96.50	35.30	2.70 ↑
Y	Y	Y	Y	Y	**100.92**	95.50	**46.00**	**13.40** ↑

augmentation, loss function adjustments, and gradient control techniques. The results of these experiments reveal important trade-offs between adversarial robustness and standard accuracy.

• **SAR** has proven effective in various domains. Initially proposed by Miyato et al. [43] for semi-supervised learning, it was later integrated into the TRADES framework for adversarial training by Zhang et al. [44]. More recently, its utility has been validated in fine-tuning tasks within natural language processing (NLP) [45].

• **Gradient clipping (GC)**, where the gradient norm is restricted within a value of 1, is utilized to stabilize Meta-AT training involving multiple attacks. This technique, widely adopted in long short-term memory (LSTM) networks [46] and proximal policy optimization (PPO) algorithms in reinforcement learning (RL) [47], effectively mitigates gradient explosion issues.

• **Warm-up learning rate (WLR)** of 0.1 is consistently employed as a fine-tuning technique [45].

• **Label smoothing (LS) [48]** is widely recognized in robust models for its ability to mitigate gradient masking effects. Following the Bag of Tricks [14], we tested LS with a smoothing value of 0.3 in Resnet-18.

• **Mixup [49]** linearly interpolates between two different training examples to create new ones, improving model robustness and reducing overfitting, which is a widely used data augmentation [50].

Table 5 illustrates that the majority of training techniques enhance AA and improve $EDSR$, albeit at a slight cost to SA. The presented SA^* and AA values were obtained using the "standard" mode of AA. The first row, denoted by 'N' (No) for all techniques, serves as the baseline for comparison. An upward arrow (↑) represents the numerical

increase relative to this baseline. In particular, SAR and GC improve the AA by 3.90% and 5.10%, respectively. WLR also demonstrates its robust fine-tuning capabilities by providing a 3.00% improvement in AA. In contrast, LS shows a marginal gain of only 0.10%. Among data augmentation methods, Mixup proves to be well-suited for Meta-AT, yielding a 2.70% increase. Finally, by incorporating all these beneficial techniques marked as 'Y' (Yes), Meta-AT achieves a peak $EDSR$ of 100.92% and an AA of 46.00%. Further techniques will be investigated in future studies.

Meta-AT maintains a highly accurate, robust, and widely applicable learning algorithm against unknown attacks. Although its AA is slightly lower than the traditional R-AT, it is still stronger than other robust defense methods and more suitable for a wider range of attacks. This shows the trade-off between defending against strong attacks and generalizing to unknown attacks. In conclusion, our MAD benchmark can be considered relatively successful.

5 Conclusion

In this paper, we introduce a novel MAD benchmark consisting of three extensive MAD datasets, an MAD evaluation toolkit, and a baseline algorithm called Meta-AT. The MAD datasets are created by subjecting the MNIST, CIFAR-10, and Tiny-ImageNet datasets to 30 mainstream adversarial attacks. In the evaluation toolkit, besides introducing a protocol about dataset configurations and the backbone network training scheme, we have also innovatively proposed a versatile evaluation metric called $EDSR$, which provides a comprehensive and unbiased assessment of the robust learning ability of defense methods. Meta-AT exhibits the capability of predictive defense. It investigates the generalization of adversarial defense methods against unknown attacks when combining online AT and offline AT, through learning from a limited number of AEs. Experimental results demonstrate that Meta-AT-trained models achieve competitive $EDSR$ values compared to other SOTA methods against both known and unknown attacks, leveraging few-shot learning to adapt effectively within a few minutes. Furthermore, these models maintain a high SA across a wide range of attack types, highlighting their robust performance and adaptability in diverse adversarial scenarios. These performance levels are comparable to those of industrial-grade model defense applications, which is a significant accomplishment. However, one limitation is that the MAD datasets created for Meta-AT are quite large, and future work will focus on reducing their size while maintaining comparable learning effectiveness. Furthermore, we also plan to optimize the Meta-AT framework and explore its application to other tasks such as object tracking [51] and large language models (LLMs) [52].

Acknowledgements This work was supported by Joint Funds of the National Natural Science Foundation of China (Grant No. U23A20346) and National Natural Science Foundation of China (Grant No. 62173107).

Supporting information Appendixes A and B. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Chen H, Yuan K, Huang Y J, et al. Feedback is all you need: from ChatGPT to autonomous driving. Sci China Inf Sci, 2023, 66: 166201
2 Wu X, Tao R, Hong D F, et al. The FrFT convolutional face: toward robust face recognition using the fractional Fourier transform and convolutional neural networks. Sci China Inf Sci, 2020, 63: 119103
3 Wang Z Q, Du Y, Wei K J, et al. Vision, application scenarios, and key technology trends for 6G mobile communications. Sci China Inf Sci, 2022, 65: 151301
4 Akhtar N, Mian A, Kardan N, et al. Advances in adversarial attacks and defenses in computer vision: a survey. IEEE Access, 2021, 9: 155161
5 Tao G, Ma S, Liu Y, et al. Attacks meet interpretability: attribute-steered detection of adversarial examples. In: Proceedings of Advances in Neural Information Processing Systems, 2018
6 Li S, Zhu S, Paul S, et al. Connecting the dots: detecting adversarial perturbations using context inconsistency. In: Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, 2020. 396–413
7 Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations. 2017. ArXiv:1711.00117
8 Raff E, Sylvester J, Forsyth S, et al. Barrage of random transforms for adversarially robust defense. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 6528–6537
9 Zhai R, Dan C, He D, et al. Macer: attack-free and scalable robust training via maximizing certified radius. 2020. ArXiv:2001.02378
10 Jia J, Cao X, Wang B, et al. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. 2019. ArXiv:1912.09899
11 Cemgil T, Ghaisas S, Dvijotham K, et al. Adversarially robust representations with smooth encoders. In: Proceedings of the International Conference on Learning Representations (ICLR), 2019
12 He Z, Rakin A S, Li J, et al. Defending and harnessing the bit-flip based adversarial weight attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 14095–14103
13 Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. 2017. ArXiv:1706.06083
14 Pang T, Yang X, Dong Y, et al. Bag of tricks for adversarial training. 2020. ArXiv:2010.00467
15 Peng S Y, Xu W, Cornelius C, et al. Robust principles: architectural design principles for adversarially robust CNNs. 2023. ArXiv:2308.16258
16 Tang S, Gong R, Wang Y, et al. Robustart: benchmarking robustness on architecture design and training techniques. 2021. ArXiv:2109.05211
17 Liu C, Dong Y, Xiang W, et al. A comprehensive study on robustness of image classification models: benchmarking and rethinking. 2023. ArXiv:2302.14301

18 Liu G R, Zhang W Z, Li X J, et al. VulnerGAN: a backdoor attack through vulnerability amplification against machine learning-based network intrusion detection systems. Sci China Inf Sci, 2022, 65: 170303
19 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324
20 Yang L, Jiang H, Cai R, et al. CondenseNetV2: sparse feature reactivation for deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 3569–3578
21 Ye M, Huang W K, Shi Z K, et al. Revisiting federated learning with label skew: an over-confidence perspective. Sci China Inf Sci, 2025, 68: 192102
22 Wong E, Rice L, Kolter J Z. Fast is better than free: revisiting adversarial training. 2020. ArXiv:2001.03994
23 Shafahi A, Najibi M, Ghiasi M A, et al. Adversarial training for free! In: Proceedings of Advances in Neural Information Processing Systems, 2019
24 Zheng H, Zhang Z, Gu J, et al. Efficient adversarial training with transferable adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 1181–1190
25 Zhang D, Zhang T, Lu Y, et al. You only propagate once: accelerating adversarial training via maximal principle. In: Proceedings of Advances in Neural Information Processing Systems, 2019
26 Farnia F, Zhang J M, Tse D. Generalizable adversarial training via spectral normalization. 2018. ArXiv:1811.07457
27 Grabinski J, Jung S, Keuper J, et al. FrequencyLowCut pooling–plug & play against catastrophic overfitting. In: Proceedings of the European Conference on Computer Vision (ECCV), 2022. 36-57
28 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the International Conference on Machine Learning (ICML), 2017. 1126–1135
29 Goldblum M, Fowl L, Goldstein T. Adversarially robust few-shot learning: a meta-learning approach. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 33: 17886–17895
30 Liu F, Zhao S, Dai X, et al. Long-term cross adversarial training: a robust meta-learning method for few-shot classification tasks. 2021. ArXiv:2106.12900
31 Qi J, Zhang R, Li C, et al. Cross domain few-shot learning via meta adversarial training. 2022. ArXiv:2202.05713
32 Ma C, Zhao C, Shi H, et al. Metaadvdet: towards robust detection of evolving adversarial attacks. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019. 692–701
33 Metzen J H, Finnie N, Hutmacher R. Meta adversarial training against universal patches. 2021. ArXiv:2101.11453
34 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770–778
35 Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), 2019. 6105–6114
36 Nicolae M I, Sinn M, Tran M N, et al. Adversarial robustness toolbox v1.0.0. 2018. ArXiv:1807.01069
37 Ding G W, Wang L, Jin X. AdverTorch v0.1: an adversarial robustness toolbox based on PyTorch. 2019. ArXiv:1902.07623
38 Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: Proceedings of the International Conference on Machine Learning (ICML), 2020. 2206–2216
39 Rice L, Wong E, Kolter J Z. Overfitting in adversarially robust deep learning. In: Proceedings of the International Conference on Machine Learning (ICML), 2020. 8093–8104
40 Rusu A A, Rabinowitz N C, Desjardins G, et al. Progressive neural networks. 2016. ArXiv:1606.04671
41 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2014. ArXiv:1412.6572
42 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems (NeurIPS), 2012. 1097–1105
43 Miyato T, Maeda S I, Koyama M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans Pattern Anal Mach Intell, 2018, 41: 1979–1993
44 Zhang H, Yu Y, Jiao J, et al. Theoretically principled trade-off between robustness and accuracy. In: Proceedings of the International Conference on Machine Learning (ICML), 2019. 7472–7482
45 Jiang H, He P, Chen W, et al. Smart: robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. 2019. ArXiv:1911.03437
46 Xie G, Shangguan A Q, Fei R, et al. Motion trajectory prediction based on a CNN-LSTM sequential model. Sci China Inf Sci, 2020, 63: 212207
47 You Q B, Ying C Y, Zhou X N, et al. Understanding adversarial attacks on observations in deep reinforcement learning. Sci China Inf Sci, 2024, 67: 152104
48 Carlini N, Athalye A, Papernot N, et al. On evaluating adversarial robustness. 2019. ArXiv:1902.06705
49 Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization. 2017. ArXiv:1710.09412
50 DeVries T, Taylor G W. Improved regularization of convolutional neural networks with cutout. 2017. ArXiv:1708.04552
51 Zha J, Fan Y, Li K, et al. Decoupled multi-hierarchy Kalman filter for 3D object tracking. 2025. ArXiv:2505.12340
52 Yang L, Zheng Z, Chen B, et al. Nullu: mitigating object hallucinations in large vision-language models via HalluSpace projection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025. 14635–14645