Fast and Reliable Evaluation of Adversarial Robustness with Minimum-Margin Attack

Ruize Gao [1] **Jiongxiao Wang** [1] **Kaiwen Zhou** [1] **Feng Liu** [2] **Binghui Xie** [1] **Gang Niu** [3]
Bo Han [4] **James Cheng** [1]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, HKSAR, China

[2]School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia

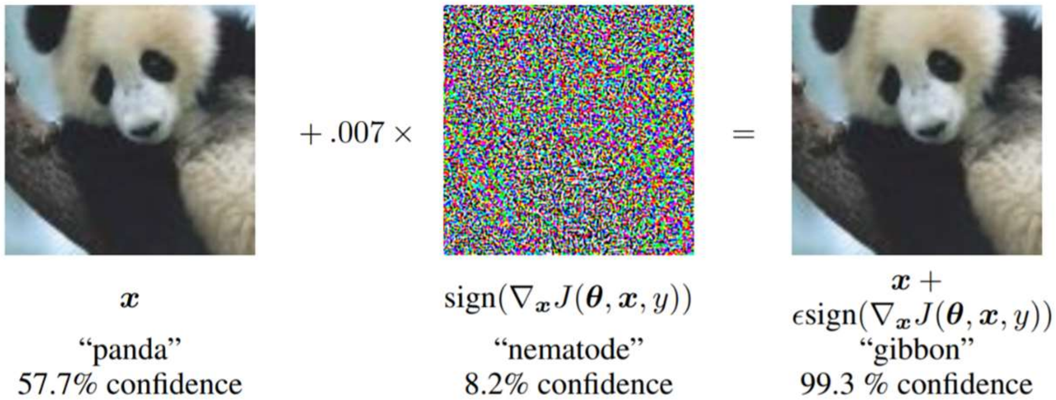
[3]RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

[4]Department of Computer Science, Hong Kong Baptist University, HKSAR, China

Background

R. Gao, et al., ICML'22

The Deep Neural Networks are Vulnerable to Adversarial Examples



From: Explaining and Harnessing Adversarial Examples. In *ICLR*, 2015.

It's necessary to find a reliable way to evaluate adversarial robustness of a DNN.

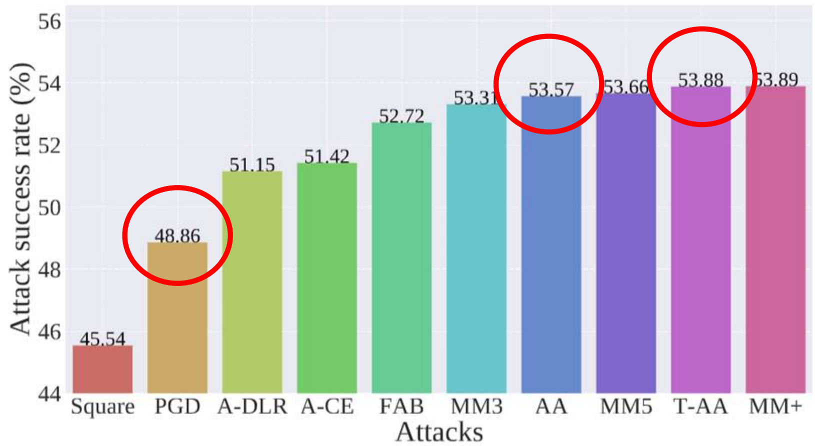
Motivations

R. Gao, et al., ICML'22

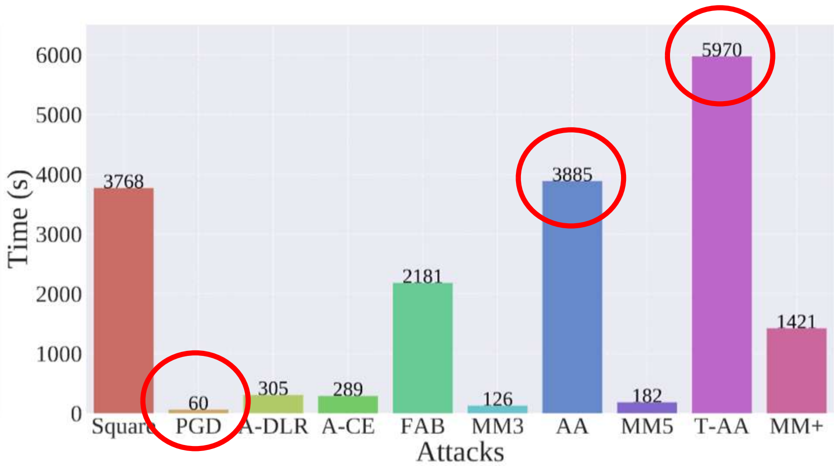
Adversarial Attack: The Dilemma between Reliability and Computational Efficiency.

Benchmark 1: Projected Gradient Descent Attack (PGD), high computational efficiency but low reliability

Benchmark 2: The attack ensemble AutoAttack, high reliability but low computational efficiency



(a) Reliability ranking



(b) Computational time

Note : MM3, MM5 and MM+ are different versions of our provided MM attack.

Minimum-Margin Attack

R. Gao, et al., ICML'22

The necessary and sufficient condition to the complete robustness of the classifier.

Condition 1. *Given a natural example x with its true label y, the K-class classifier f satisfies*

$$\forall x' \in \mathcal{B}_\epsilon[x], z_y(x') - \max_{i \neq y} z_i(x') \geq 0,$$

where $\mathcal{B}_\epsilon[x] = \{x' \mid d_\infty(x, x') \leq \epsilon\}$; $z_y(x') = f(x')_y$; $z_i(x') = f(x')_i$.

According the **condition 1**, we define the most adversarial example.

Definition 1 (The most adversarial example)**.** *Given a natural example x with its true label y, the most adversarial example x^* within $\mathcal{B}_\epsilon[x]$ is defined as:*

$$\forall x' \in \mathcal{B}_\epsilon[x], x^* = \arg\max_{x'} -(z_y(x') - \max_{i \neq y} z_i(x')),$$

where $\mathcal{B}_\epsilon[x] = \{x' \mid d_\infty(x, x') \leq \epsilon\}$ *is the closed ball of radius* $\epsilon > 0$ *centered at* x; $z_y(x') = f(x')_y$; $z_i(x') = f(x')_i$.

Minimum-Margin Attack

R. Gao, et al., ICML'22

Using margin to identify the "most adversarial example"

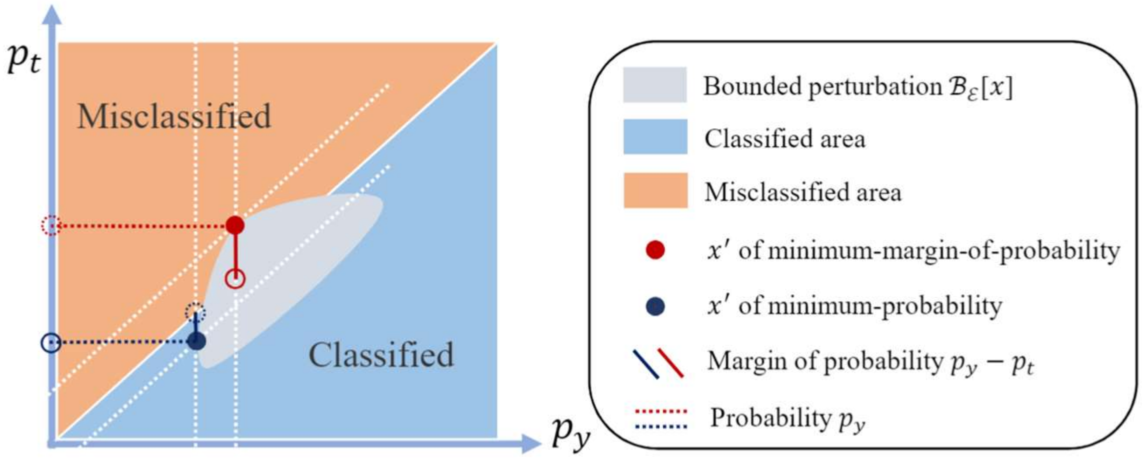


Figure 2. Minimum margin of probability. p denotes the predicted probability, p_y and p_t are the predicted probability on the true label y and a targeted false label t. The gray shape is the image of the adversarial variants x' within the bounded perturbation ball $\mathcal{B}_\epsilon[x]$ under the mapping of the network onto (p_y, p_t); the orange area ($p_t > p_y$) indicates the region where the adversarial variants are misclassified, or to say a successful attack, while the blue area ($p_t < p_y$) indicates the region where the adversarial variants do not attack successfully.

Minimum-Margin Attack

R. Gao, et al., ICML'22

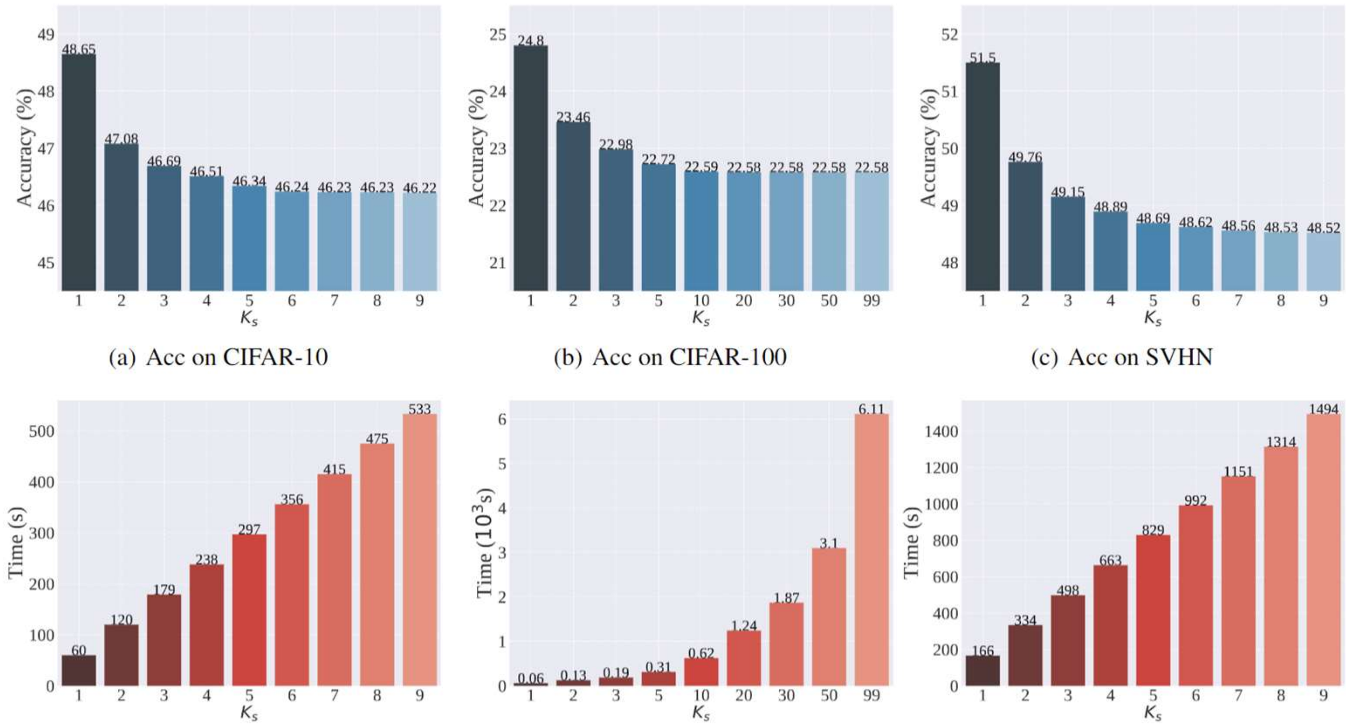
Sequential TArget Ranking Selection (STATS)

1)Pre-selecting-Targets Strategy:

Selecting partial targets achieves comparable performance.

2)Ranking-Sequential-Attack Strategy:

Consider the false target with the highest predicted probability first; if the attack succeeds, then terminate attacks on other targets; otherwise, continue considering the false target with the second highest predicted probability.



(a) Acc on CIFAR-10 (b) Acc on CIFAR-100 (c) Acc on SVHN

(d) Time on CIFAR-10 (e) Time on CIFAR-100 (f) Time on SVHN

Minimum-Margin Attack

R. Gao, et al., ICML'22

With the mentioned strategies, we summarize our scheme of MM attack.

Condition 2 and Condition 3 follow the setting of the adaptive step size selection in [1]:

Condition 2. $\sum_{i=w_{j-1}}^{w_j-1} 1_{f(x'_{i+1})>f(x'_i)} < \beta \cdot (w_j - w_{j-1}).$

Condition 3. $\alpha^{w_{j-1}} \equiv \alpha^{w_j}$ and $f_{max}^{w_{j-1}} \equiv f_{max}^{w_j}.$

Algorithm 1 MM Attack

Input: natural data x, true label y, set of false labels C, model f, loss function ℓ_{MM}, maximum number of PGD steps N, perturbation bound ϵ, initial step size α, the number of classes K, targets selection number K_s, checkpoints set W;
Output: adversarial data x';
while $K_s > 0$ **do**
  $x'_0 \leftarrow x$;
  $x'_{max} \leftarrow x$;
  $f_{max} \leftarrow f(x'_0)$;
  $c = \arg\max_{i \in C} f(x)_i$;
  for $k = 0$ **to** $N-1$ **do**
    $x'_{k+1} \leftarrow \Pi_{\mathcal{B}_\epsilon[x]}(x'_k + \alpha sign(\nabla_{x'_k} \ell_{MM}(f(x'_k), y, c))$;
    if $f(x'_{k+1}) > f_{max}$ **then**
      $x'_{max} \leftarrow x'_{k+1}$;
      $f_{max} \leftarrow f(x'_{k+1})$;
    end if
    if $k \in W$ **and** (Condition 2 **or** Condition 3) **then**
      $\alpha \leftarrow \alpha/2$;
      $x'_{k+1} \leftarrow x'_{max}$;
    end if
  end for
  $C \leftarrow C \setminus \{c\}$;
  if $\arg\max_{i \in C} f(x')_i \neq y$ **then**
    $K_s \leftarrow 0$;
  end if
  $K_s \leftarrow K_s - 1$;
end while

Reference:
[1]:Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In ICML, 2020.

Experiments

Baselines:

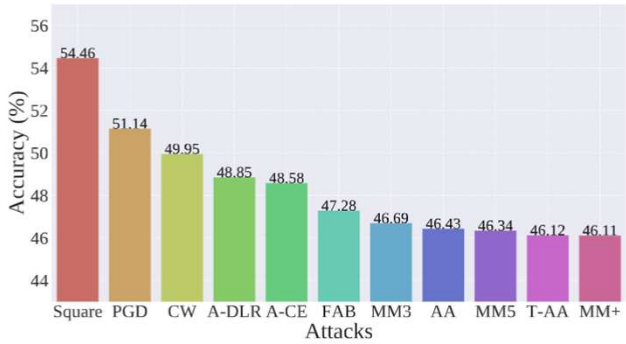
- PGD: Projected Gradient Descent Attack [1]
- CW: Carlini and Wagner attack [2]
- A-DLR: PGD with adaptive step size and DLR loss [3]
- A-CE: PGD with adaptive step size and CE loss [3]
- FAB: A component of the AutoAttack ensemble [3]
- Square: A component of the AutoAttack ensemble [3]
- AA: AutoAttack with untargeted version [3]
- T-AA: AutoAttack with targeted version [3]

Reference:
[1]:Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant toadversarial attacks. In ICLR, 2018.
[2]:Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In CVPR, 2017.
[3]:Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In ICML, 2020.

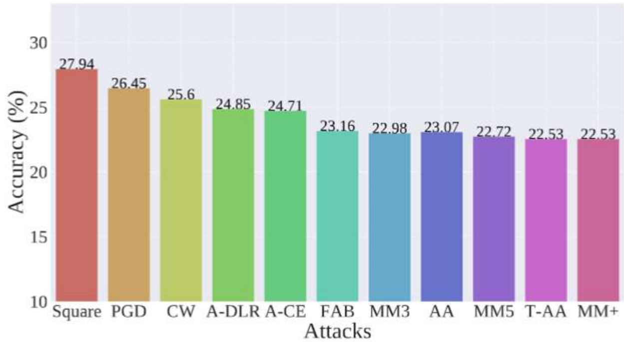
Experiments

R. Gao, et al., ICML'22

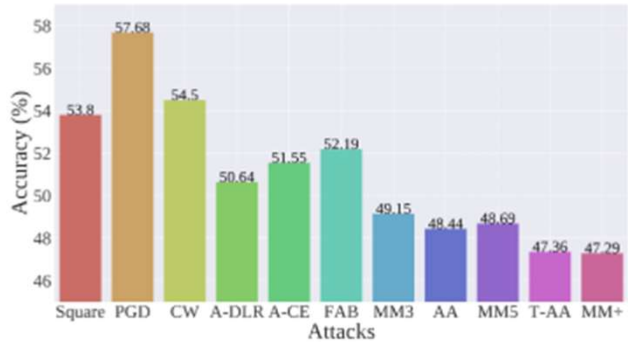
Main results on different datasets.



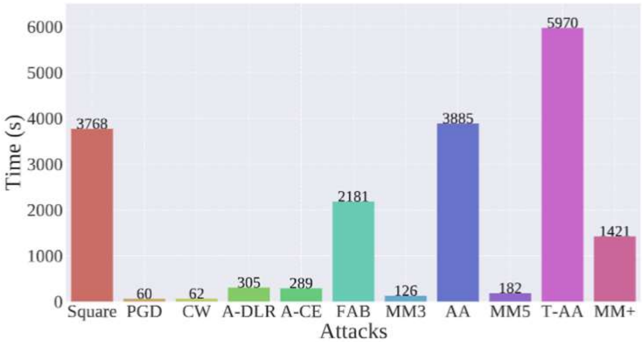
(a) Evaluation on CIFAR-10



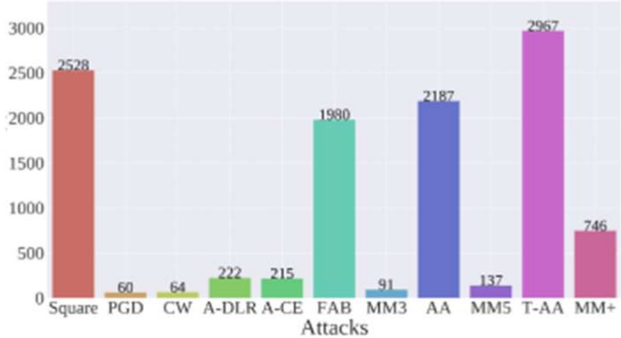
(c) Evaluation on CIFAR-100



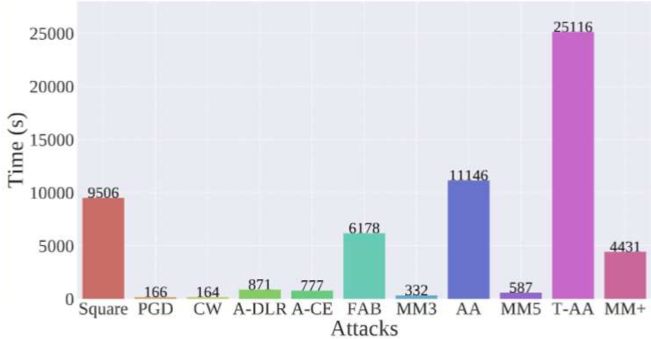
(e) Evaluation on SVHN



(b) Computational time on CIFAR-10



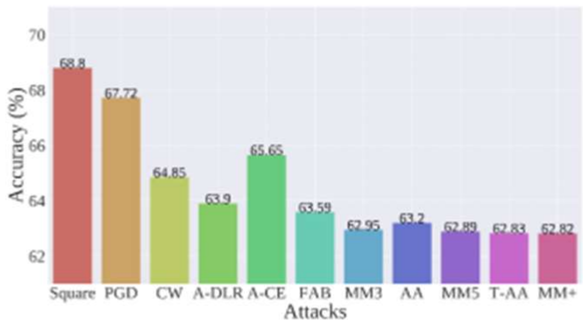
(d) Computational time on CIFAR-100



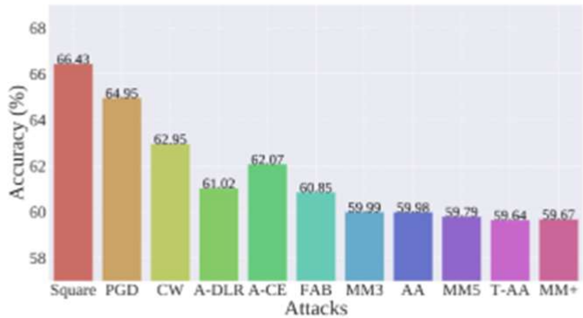
(f) Computational time on SVHN

Experiments

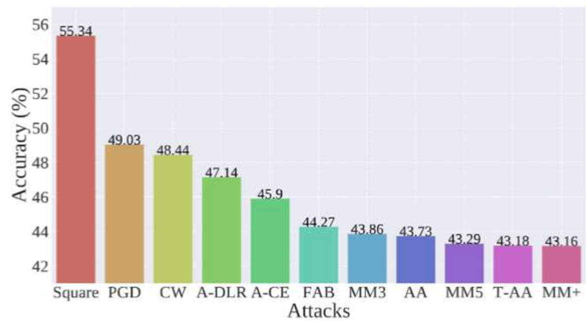
Main results on different well-trained models provided in RobustBench.



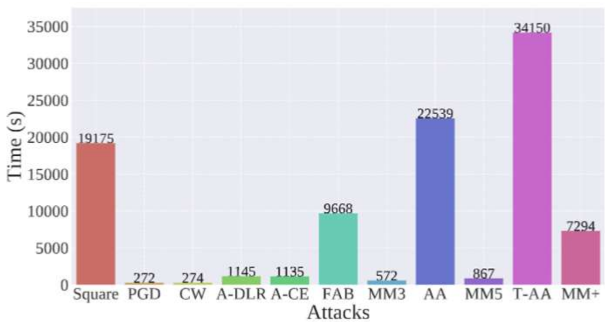
(a) Evaluation on Gowal et al. (2020)



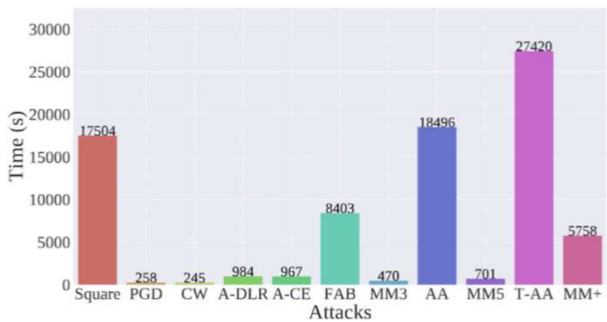
(c) Evaluation on Sridhar et al. (2021)



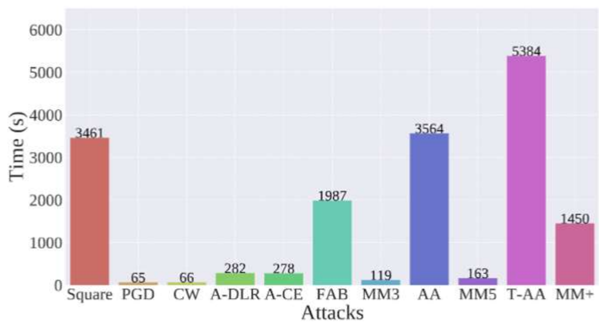
(e) Evaluation on Wong et al. (2020)



(b) Computational time on Gowal et al. (2020)



(d) Computational time on Sridhar et al. (2021)



(f) Computational time on Wong et al. (2020)

Experiments

R. Gao, et al., ICML'22

Adversarial Training with MM Attack.

Main results on *CIFAR-10*

Methods	PGD	Diff.	CW	Diff.	MM3-F10	Diff.	MM3-F20	Diff.	MM3	Diff.
PGD (Test)	51.14	-4.10	51.47	-3.77	54.96	-0.28	**55.24**	0.00	55.04	-0.20
CW (Test)	49.95	-1.89	**53.26**	0.00	51.18	-2.08	51.16	-2.10	51.84	-1.42
A-CE (Test)	48.58	-3.92	48.16	-4.34	51.55	-0.95	**52.50**	0.00	52.22	-0.28
A-DLR (Test)	48.85	-1.44	**52.76**	0.00	49.78	-2.98	49.88	-2.88	50.29	-2.47
FAB (Test)	47.28	-1.22	47.13	-1.37	47.83	-0.67	48.28	-0.22	**48.50**	-0.00
Square (Test)	54.46	-0.66	**55.32**	0.00	54.80	-0.52	54.83	-0.49	55.12	-0.20
AA (Test)	46.43	-1.85	46.36	-1.92	47.62	-0.66	47.84	-0.44	**48.28**	-0.00
T-AA (Test)	46.12	-0.97	45.26	-1.83	46.39	-0.70	46.73	-0.36	**47.09**	-0.00
MM3 (Test)	46.69	-1.17	46.77	-1.09	47.20	-0.66	47.48	-0.38	**47.86**	-0.00
MM9 (Test)	46.21	-0.95	45.36	-1.80	46.49	-0.67	46.82	-0.34	**47.16**	-0.00
MM+ (Test)	46.12	-0.90	45.22	-1.80	46.39	-0.63	46.68	-0.34	**47.02**	-0.00

Acknowledgements

RZG, JXW, KWZ, BHX and JC were supported by GRF 14208318 from the RGC of HKSAR. BH was supported by the RGC Early Career Scheme No. 22200720, NSFC Young Scientists Fund No. 62006202, and Guangdong Basic and Applied Basic Research Foundation No. 2022A1515011652. GN were supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, Japan.

Thank You!