Simulating Unknown Target Models for Query-Efficient Black-box Attacks Supplementary Material

Chen Ma, Li Chen*, and Jun-Hai Yong
School of Software, BNRist, Tsinghua University, Beijing, China
mac16@mails.tsinghua.edu.cn, {chenlee,yongjh}@tsinghua.edu.cn

1. Experiment Settings

1.1. Compared Methods

Bandits. Table 1 shows the default hyperparameters of Bandits [5], which is a subset of hyperparameters of the Simulator Attack. Specifically, the OCO learning rate is used to update the prior, which is an alias of the gradient $\mathbf{g}$ for updating the input image.

RGF and P-RGF. Table 2 shows the default hyperparameters of random gradient-free (RGF) [11] and prior-guided RGF (P-RGF) [1]. P-RGF improves RGF by using surrogate models (see the last row block of Table 2). The experiments of RGF and P-RGF are conducted by using the implementation of PyTorch version that is translated from the official TensorFlow version.

NES. The default hyperparameters for natural evolution strategies (NES) [4] are listed in Table 5. In the targeted attack, NES uses an initial image of the target class and reduce its distortion iteratively while keeping the image residing in the adversarial region of the target class. Finally, the samples whose ℓ_p norm distance to the original benign image is less than a preset ϵ are considered as successful samples. Thus, the hyperparameters of NES are carefully tuned in the untargeted and targeted attack separately, so as to achieve the highest attack success rate. The experiments of NES are conducted by using the implementation of PyTorch version, which is translated from the official TensorFlow implementation.

Meta Attack. The default hyperparameters of the Meta Attack [2] are listed in Table 6. Specifically, the meta interval m is set to 3 in two cases, namely, the targeted attack and all the experiments of TinyImageNet dataset. In other cases, the meta interval m is set to 5. The gradients of training data are generated by using the classification networks listed in Table 7. The Meta Attack uses the official PyTorch implementation to conduct ℓ_2 norm attack experiments, and we add the additional code in the official implementation to enable it to support the ℓ_∞ norm attack.

*Corresponding author.

Norm	Hyperparameter	Value
ℓ_2	δ, finite difference probe	0.01
	η, image learning rate	0.1
	η_g, OCO learning rate	0.1
	τ, Bandits exploration	0.3
	ϵ, radius of ℓ_2 norm ball	4.6
	maximum query times	10,000
ℓ_∞	δ, finite difference probe	0.1
	η, image learning rate	1/255
	η_g, OCO learning rate	1.0
	τ, Bandits exploration	0.3
	ϵ, radius of ℓ_∞ norm ball	8/255
	maximum query times	10,000

Table 1: The hyperparameters of Bandits [5].

Norm	Hyperparameter	Value
ℓ_2	h, image learning rate	2.0
	σ, sampling variance	1e-4
	ϵ, radius of ℓ_2 norm ball	4.6
ℓ_∞	h, image learning rate	0.005
	σ, sampling variance	1e-4
	ϵ, radius of ℓ_∞ norm ball	8/255
ℓ_2, ℓ_∞	surrogate model used in CIFAR-10/100	ResNet-110
ℓ_2, ℓ_∞	surrogate model used in TinyImageNet	ResNet-101

Table 2: The hyperparameters of RGF [11] and P-RGF [1], and the networks shown in the last row block are used as the surrogate models of P-RGF.

Simulator Attack. The default hyperparameters of the proposed method are listed in Table 3. Those hyperparameters that are also used in Bandits are set to the same values as Bandits.

1.2. Pre-trained Networks and Target Models

Pre-trained Networks. In the training of the Simulator and the auto-encoder of the Meta Attack, we collect various types of classification networks to generate the training data. In our experiments, we select 14 networks for generating training data of CIFAR-10 and CIFAR-100 datasets, and select 16 networks for generating training data of TinyImageNet datasets. The names of these networks and their training configurations are shown in Table 7.

Hyperparameter	Default Value
backbone	ResNet-34
λ_1, the learning rate of the inner update	0.01
λ_2, the learning rate of of the outer update	0.001
ϵ, the maximum distortion of ℓ_2 norm attack	4.6
ϵ, the maximum distortion of ℓ_∞ norm attack	8/255
δ, finite difference probe of ℓ_2 norm attack	0.01
δ, finite difference probe of ℓ_∞ norm attack	0.1
η, the image learning rate of ℓ_2 norm attack	0.1
η, the image learning rate of ℓ_∞ norm attack	1/255
η_g, OCO learning rate of ℓ_2 norm attack	0.1
η_g, OCO learning rate of ℓ_∞ norm attack	1.0
τ, Bandits exploration	0.3
inner-update iterations	12
meta-predict interval m	5
warm-up iterations t	10
deque $\mathbb{D}$'s maximum length	10

Table 3: The hyperparameters of the Simulator Attack.

Dataset	Network	Model Details		
		Params(M)	MACs(G)	Layers
CIFAR-10	PyramidNet-272	26.21	4.55	272
	GDAS	3.02	0.41	20
	WRN-28	36.48	5.25	28
	WRN-40	55.84	8.08	40
CIFAR-100	PyramidNet-272	26.29	4.55	272
	GDAS	3.14	0.41	20
	WRN-28	36.54	5.25	28
	WRN-40	55.90	8.08	40
TinyImageNet	DenseNet-121	7.16	0.23	121
	ResNeXt-101 (32×4d)	42.54	0.65	101
	ResNeXt-101 (64×4d)	81.82	1.27	101

Table 4: The details of black-box target models which are used for evaluating attack methods, where MAC is the multiply–accumulate operation count.

Target Models. To evaluate the performance of attacking *unknown* target models, we specify the target models to equip with completely different architectures from the pre-trained networks. The target models and their complexity are listed in Table 4.

2. Experimental Results

2.1. Detailed Experimental Figures

Attack Success Rates at Different Maximum Queries. We conduct experiments by limiting different maximum queries of attacks and compare their attack success rates. Figs. 1, 2, 3, and 4 show the results which are obtained by attacking normal models and defensive models with different maximum number of queries. Four defensive models are adopted, namely, ComDefend [6], Feature Distillation [8], prototype conformity loss (PCL) [10] and Adv Train [9]. The ResNet-50 [3] is selected as the backbone in these defensive models.

Average Queries at Different Success Rates. The second type of figure measures the average number of queries that reaches different desired success rates. It demonstrates the relation between the query number and attack success rate from a different angle. Specifically, given a desired success rate a and the query list Q of all successful attacked samples, the average query (Avg. Q_a) is defined as follows:

$$\text{Avg. } Q_a = \frac{\sum_{i=1}^{N} \hat{Q}_i}{N}, \qquad \text{where } \hat{Q} = Q[Q \leq P_a], \quad (1)$$

where P_a is the a-th percentile value of Q and N is the length of $\hat{Q}$. Figs. 5, 6, 7, and 8 show the results. All the

Dataset	Attack	Norm	Hyperparameter	Value
CIFAR-10	Untargeted	ℓ_2	ϵ, radius of ℓ_2 norm ball	4.6
			h, image learning rate	2.0
		ℓ_∞	ϵ, radius of ℓ_∞ norm ball	8/255
			h, image learning rate	1e-2
	Targeted	ℓ_2	ϵ_0, initial distance from the source image	20.0
			ϵ, final radius of ℓ_2 norm ball	4.6
			δ_{ϵ_0}, initial rate of decaying ϵ	1.0
			$\delta_{\epsilon_{\min}}$, the minimum rate of decaying ϵ	0.1
			$h_{\max}$, the maximum image learning rate	2.0
			$h_{\min}$, the minimum image learning rate	5e-5
		ℓ_∞	ϵ_0, initial distance from the source image	1.0
			ϵ, final radius of ℓ_∞ norm ball	8/255
			δ_{ϵ_0}, initial rate of decaying ϵ	0.1
			$\delta_{\epsilon_{\min}}$, the minimum rate of decaying ϵ	0.01
			$h_{\max}$, the maximum image learning rate	0.1
			$h_{\min}$, the minimum image learning rate	0.01
CIFAR-100	Untargeted	ℓ_2	ϵ, radius of ℓ_2 norm ball	4.6
			h, image learning rate	2.0
		ℓ_∞	ϵ, radius of ℓ_∞ norm ball	8/255
			h, image learning rate	1e-2
	Targeted	ℓ_2	ϵ_0, initial distance from the source image	20.0
			ϵ, final radius of ℓ_2 norm ball	4.6
			δ_{ϵ_0}, initial rate of decaying ϵ	1.0
			$\delta_{\epsilon_{\min}}$, the minimum rate of decaying ϵ	0.3
			$h_{\max}$, the maximum image learning rate	1.0
			$h_{\min}$, the minimum image learning rate	5e-5
		ℓ_∞	ϵ_0, initial distance from the source image	1.0
			ϵ, final radius of ℓ_∞ norm ball	8/255
			δ_{ϵ_0}, initial rate of decaying ϵ	0.1
			$\delta_{\epsilon_{\min}}$, the minimum rate of decaying ϵ	0.01
			$h_{\max}$, the maximum image learning rate	0.1
			$h_{\min}$, the minimum image learning rate	0.01
TinyImageNet	Untargeted	ℓ_2	ϵ, radius of ℓ_2 norm ball	4.6
			h, image learning rate	2.0
		ℓ_∞	ϵ, radius of ℓ_∞ norm ball	8/255
			h, image learning rate	1e-2
	Targeted	ℓ_2	ϵ_0, initial distance from the source image	40.0
			ϵ, final radius of ℓ_2 norm ball	4.6
			δ_{ϵ_0}, initial rate of decaying ϵ	1.0
			$\delta_{\epsilon_{\min}}$, the minimum rate of decaying ϵ	0.1
			$h_{\max}$, the maximum image learning rate	2.0
			$h_{\min}$, the minimum image learning rate	0.5
		ℓ_∞	ϵ_0, initial distance from the source image	1.0
			ϵ, final radius of ℓ_∞ norm ball	8/255
			δ_{ϵ_0}, initial rate of decaying ϵ	0.1
			$\delta_{\epsilon_{\min}}$, the minimum rate of decaying ϵ	1e-3
			$h_{\max}$, the maximum image learning rate	0.1
			$h_{\min}$, the minimum image learning rate	0.01

Table 5: The hyperparameters of NES [4], where the sampling variance σ for gradient estimation is set to 1e-3, and the number of samples per draw is set to 50.

Dataset	Attack	Norm	Hyperparameter	Value
CIFAR-10/100	Untargeted	ℓ_2	h, image learning rate	1e-2
			top-q coordinates for estimating gradient	125
			m, meta interval	5
			use_tanh, change-of-variables method	true
			ϵ, radius of ℓ_2 norm ball	4.6
		ℓ_∞	h, image learning rate	1e-2
			top-q coordinates for estimating gradient	125
			m, meta interval	5
			use_tanh, change-of-variables method	false
			ϵ, radius of ℓ_∞ norm ball	8/255
	Targeted	ℓ_2	h, image learning rate	1e-2
			top-q coordinates for estimating gradient	125
			m, meta interval	3
			use_tanh, change-of-variables method	true
			ϵ, radius of ℓ_2 norm ball	4.6
		ℓ_∞	h, image learning rate	1e-2
			top-q coordinates for estimating gradient	125
			m, meta interval	3
			use_tanh, change-of-variables method	false
			ϵ, radius of ℓ_∞ norm ball	8/255
TinyImageNet	Untargeted	ℓ_2	h, image learning rate	1e-2
			top-q coordinates for estimating gradient	125
			m, meta interval	3
			use_tanh, change-of-variables method	true
			ϵ, radius of ℓ_2 norm ball	4.6
		ℓ_∞	h, image learning rate	1e-2
			top-q coordinates for estimating gradient	125
			m, meta interval	3
			use_tanh, change-of-variables method	false
			ϵ, radius of ℓ_∞ norm ball	8/255
	Targeted	ℓ_2	h, image learning rate	1e-2
			top-q coordinates for estimating gradient	125
			m, meta interval	3
			use_tanh, change-of-variables method	true
			ϵ, radius of ℓ_2 norm ball	4.6
		ℓ_∞	h, image learning rate	1e-2
			top-q coordinates for estimating gradient	125
			m, meta interval	3
			use_tanh, change-of-variables method	false
			ϵ, radius of ℓ_∞ norm ball	8/255

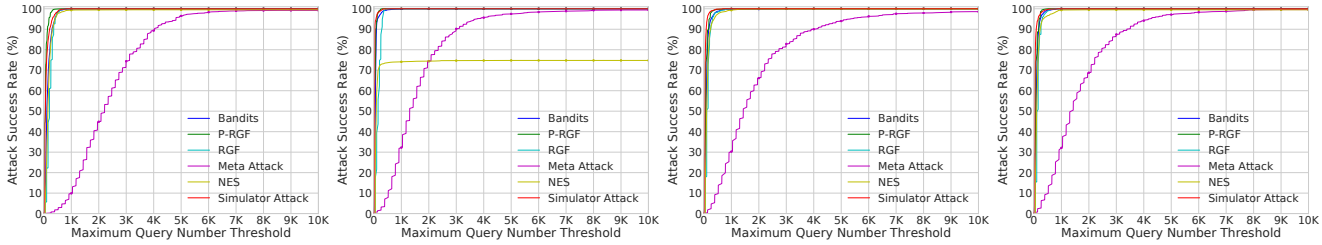
Table 6: The hyperparameters of the Meta Attack, where the binary step is set to 1, and the solver of gradient estimation adopts the Adam optimizer [7].

experimental results demonstrate that the Simulator Attack requires the lowest queries and achieves the highest attack success rate, so the superior performance of the Simulator Attack is verified.

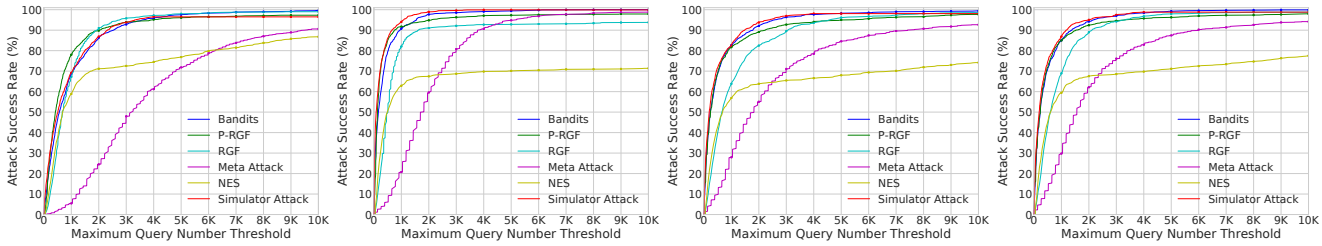
Histogram of Query Numbers. To observe the distribution of query numbers in detail, we collect the query number of each adversarial example to draw the histogram figures. Specifically, we divide the range of query number into 10 intervals, and then count the number of samples in each interval. These intervals are separated by the vertical lines of figures. Each bar indicates one attack, and its height indicates the number of samples with the queries belong to this query interval. Figs. 9, 10, 11, and 12 show the histograms of query numbers in the CIFAR-10, CIFAR-100, and TinyImageNet datasets, respectively. The results demonstrates that the highest red bars (the Simulator Attack) are located in the area with low number of queries, which confirms that most adversarial examples of the Simulator Attack have the minimum number of queries.

Dataset	Network	Training Configuration					Hyperparameters	
		epochs	lr	lr decay epochs	lr decay rate	weight decay	layer depth	other hyperparameters
CIFAR-10/100	AlexNet	164	0.1	81, 122	0.1	5e-4	9	-
	DenseNet-100	300	0.1	150, 225	0.1	1e-4	100	growth rate:12, compression rate:2
	DenseNet-190	300	0.1	150, 225	0.1	1e-4	190	growth rate:40, compression rate:2
	PreResNet-110	164	0.1	81, 122	0.1	1e-4	110	block name: BasicBlock
	ResNeXt-29 ($8 \times 64d$)	300	0.1	150, 225	0.1	5e-4	29	widen factor:4, cardinality:8
	ResNeXt-29 ($16 \times 64d$)	300	0.1	150, 225	0.1	5e-4	29	widen factor:4, cardinality:16
	VGG-19 (BN)	164	0.1	81, 122	0.1	5e-4	19	-
	ResNet-20	164	0.1	81, 122	0.1	1e-4	20	block name: BasicBlock
	ResNet-32	164	0.1	81, 122	0.1	1e-4	32	block name: BasicBlock
	ResNet-44	164	0.1	81, 122	0.1	1e-4	44	block name: BasicBlock
	ResNet-50	164	0.1	81, 122	0.1	1e-4	50	block name: BasicBlock
	ResNet-56	164	0.1	81, 122	0.1	1e-4	56	block name: BasicBlock
	ResNet-110	164	0.1	81, 122	0.1	1e-4	110	block name: BasicBlock
	ResNet-1202	164	0.1	81, 122	0.1	1e-4	1202	block name: BasicBlock
TinyImageNet	VGG-11	300	1e-3	100, 200	0.1	1e-4	11	-
	VGG-11 (BN)	300	1e-3	100, 200	0.1	1e-4	11	-
	VGG-13	300	1e-3	100, 200	0.1	1e-4	13	-
	VGG-13 (BN)	300	1e-3	100, 200	0.1	1e-4	13	-
	VGG-16	300	1e-3	100, 200	0.1	1e-4	16	-
	VGG-16 (BN)	300	1e-3	100, 200	0.1	1e-4	16	-
	VGG-19	300	1e-3	100, 200	0.1	1e-4	19	-
	VGG-19 (BN)	300	1e-3	100, 200	0.1	1e-4	19	-
	ResNet-18	300	1e-3	100, 200	0.1	1e-4	18	block name: BasicBlock
	ResNet-34	300	1e-3	100, 200	0.1	1e-4	34	block name: BasicBlock
	ResNet-50	300	1e-3	100, 200	0.1	1e-4	50	block name: Bottleneck
	ResNet-101	300	1e-3	100, 200	0.1	1e-4	101	block name: Bottleneck
	ResNet-152	300	1e-3	100, 200	0.1	1e-4	152	block name: Bottleneck
	DenseNet-161	300	1e-3	100, 200	0.1	1e-4	161	growth rate: 32
	DenseNet-169	300	1e-3	100, 200	0.1	1e-4	169	growth rate: 32
	DenseNet-201	300	1e-3	100, 200	0.1	1e-4	201	growth rate: 32

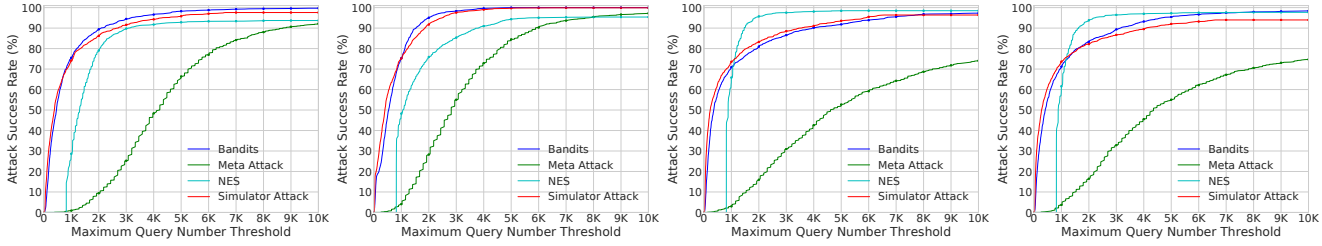
Table 7: The details of pre-trained classification networks, which are $\mathbb{N}_1$,$\cdots$,$\mathbb{N}_n$ used for the generation of training data in both the Simulator Attack and the Meta Attack. All the data of ResNet networks are excluded in the experiments of attacking defensive models.



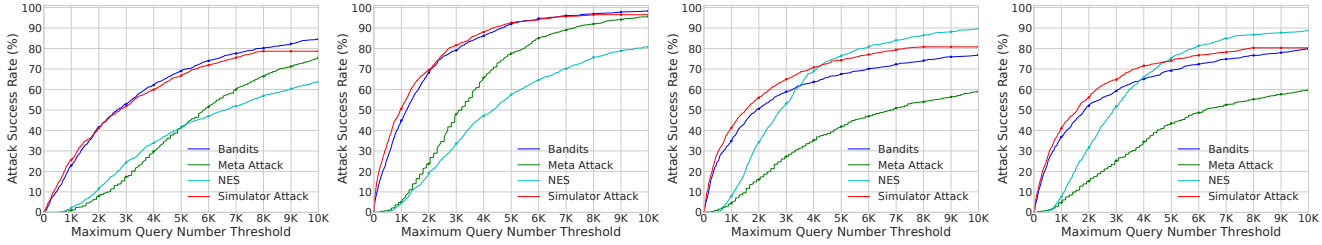
(a) untargeted ℓ_2 attack PyramidNet-272 (b) untargeted ℓ_2 attack GDAS (c) untargeted ℓ_2 attack WRN-28 (d) untargeted ℓ_2 attack WRN-40



(e) untargeted ℓ_∞ attack PyramidNet-272 (f) untargeted ℓ_∞ attack GDAS (g) untargeted ℓ_∞ attack WRN-28 (h) untargeted ℓ_∞ attack WRN-40

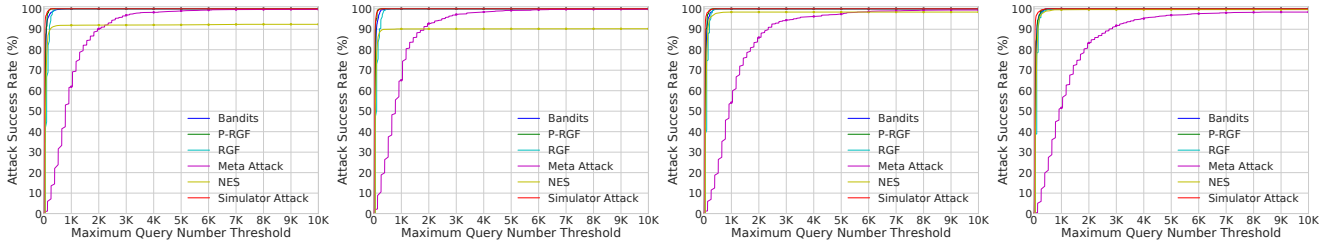


(i) targeted ℓ_2 attack PyramidNet-272 (j) targeted ℓ_2 attack GDAS (k) targeted ℓ_2 attack WRN-28 (l) targeted ℓ_2 attack WRN-40

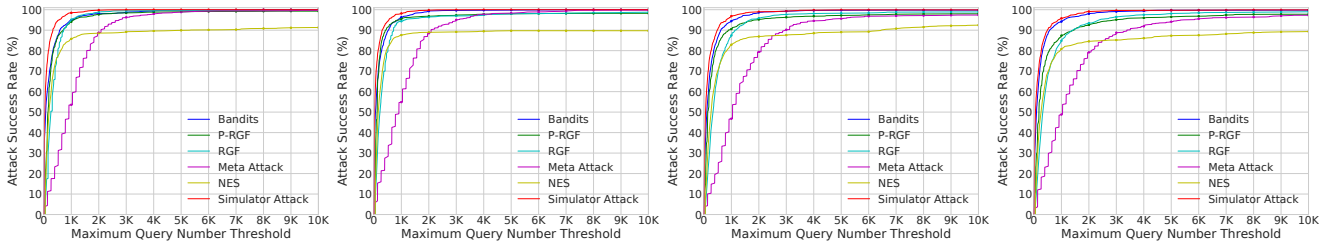


(m) targeted ℓ_∞ attack PyramidNet-272 (n) targeted ℓ_∞ attack GDAS (o) targeted ℓ_∞ attack WRN-28 (p) targeted ℓ_∞ attack WRN-40

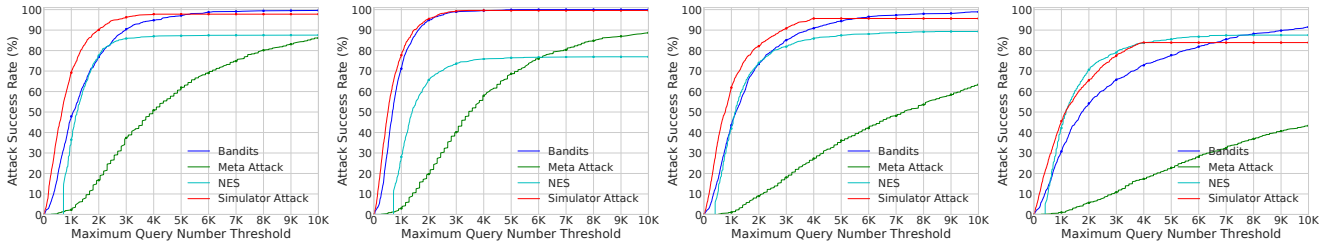
Figure 1: Comparisons of attack success rates at different limited maximum queries in CIFAR-10 dataset.



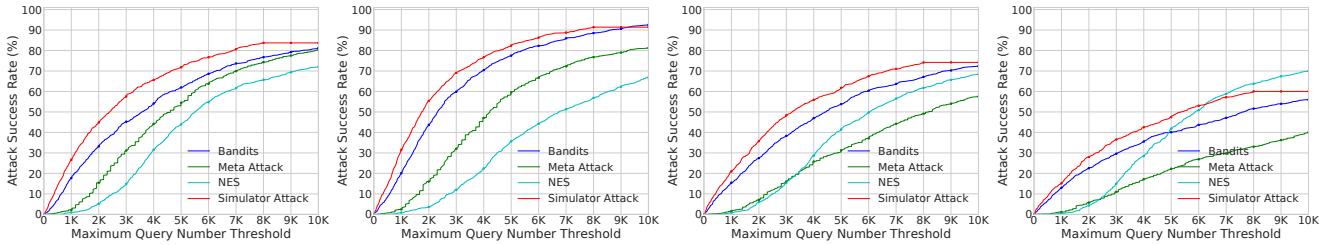
(a) untargeted ℓ_2 attack PyramidNet-272 (b) untargeted ℓ_2 attack GDAS (c) untargeted ℓ_2 attack WRN-28 (d) untargeted ℓ_2 attack WRN-40



(e) untargeted ℓ_∞ attack PyramidNet-272 (f) untargeted ℓ_∞ attack GDAS (g) untargeted ℓ_∞ attack WRN-28 (h) untargeted ℓ_∞ attack WRN-40

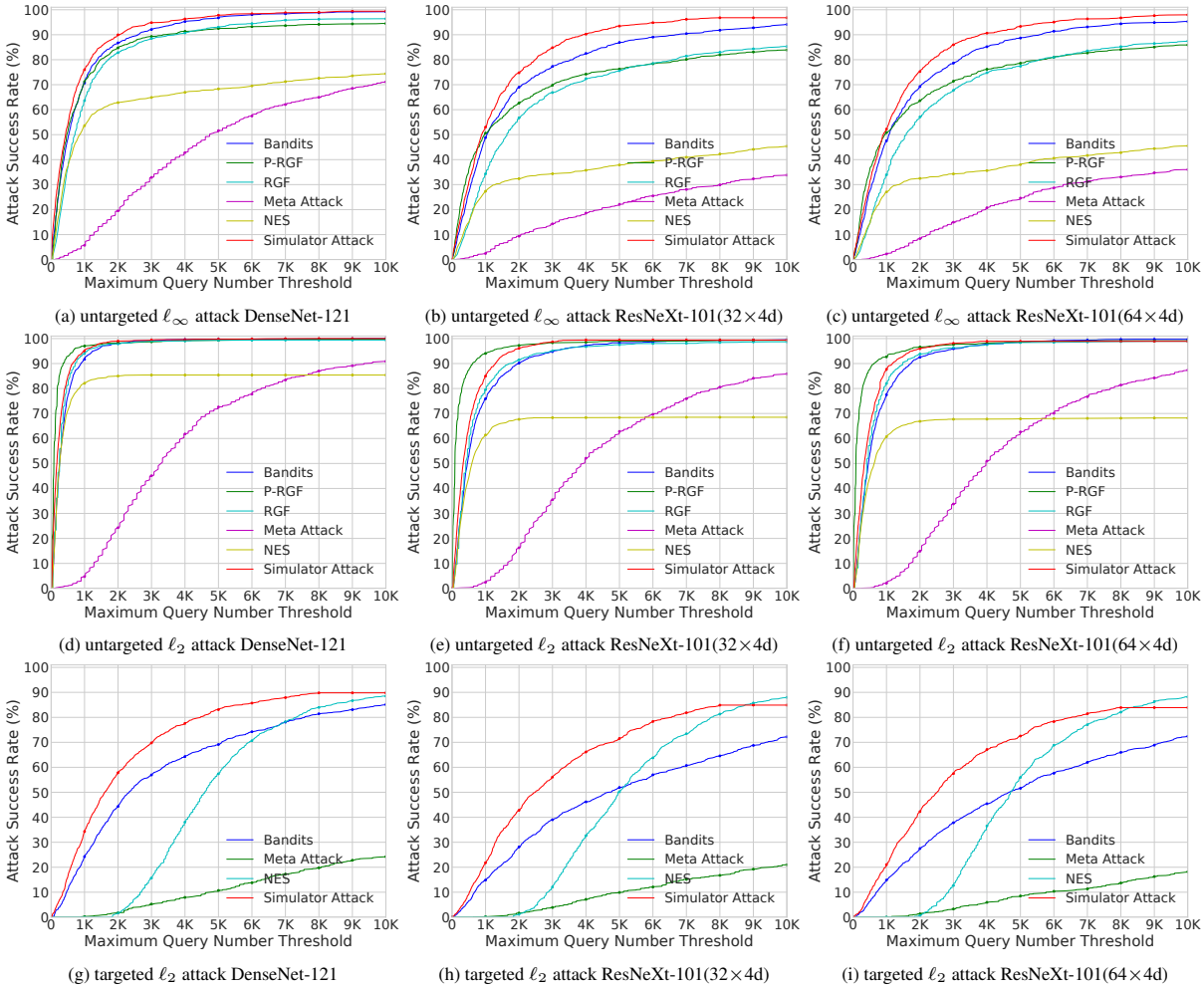


(i) targeted ℓ_2 attack PyramidNet-272 (j) targeted ℓ_2 attack GDAS (k) targeted ℓ_2 attack WRN-28 (l) targeted ℓ_2 attack WRN-40



(m) targeted ℓ_∞ attack PyramidNet-272 (n) targeted ℓ_∞ attack GDAS (o) targeted ℓ_∞ attack WRN-28 (p) targeted ℓ_∞ attack WRN-40

Figure 2: Comparisons of attack success rates at different limited maximum queries in CIFAR-100 dataset.

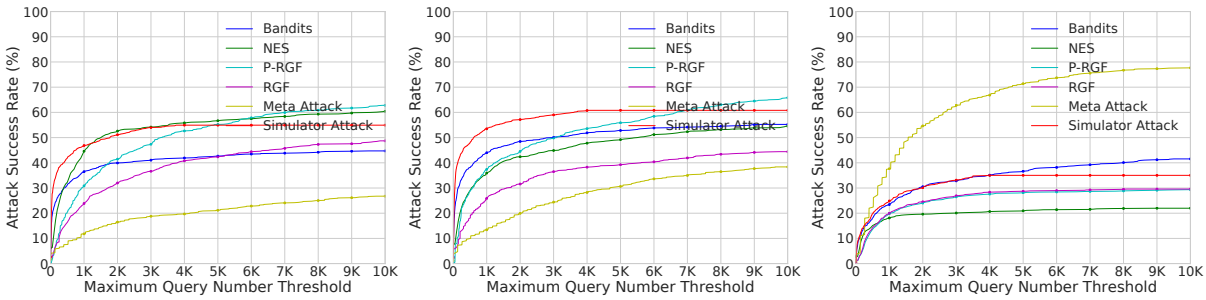


(a) untargeted ℓ_∞ attack DenseNet-121 (b) untargeted ℓ_∞ attack ResNeXt-101(32×4d) (c) untargeted ℓ_∞ attack ResNeXt-101(64×4d)

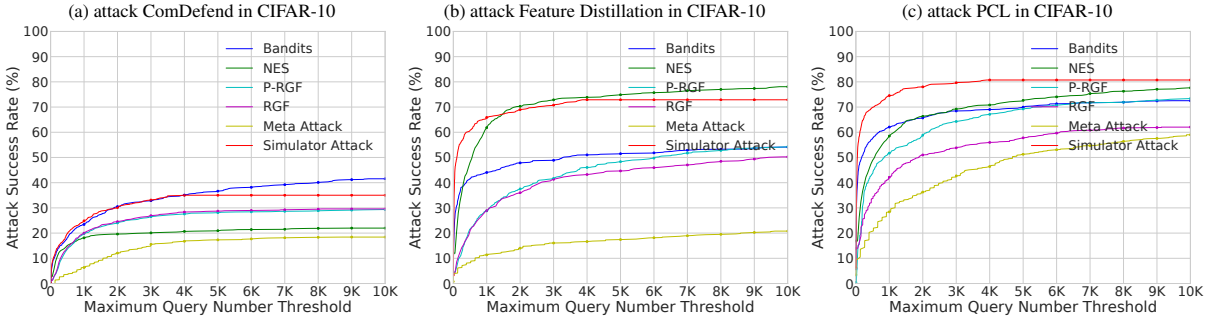
(d) untargeted ℓ_2 attack DenseNet-121 (e) untargeted ℓ_2 attack ResNeXt-101(32×4d) (f) untargeted ℓ_2 attack ResNeXt-101(64×4d)

(g) targeted ℓ_2 attack DenseNet-121 (h) targeted ℓ_2 attack ResNeXt-101(32×4d) (i) targeted ℓ_2 attack ResNeXt-101(64×4d)

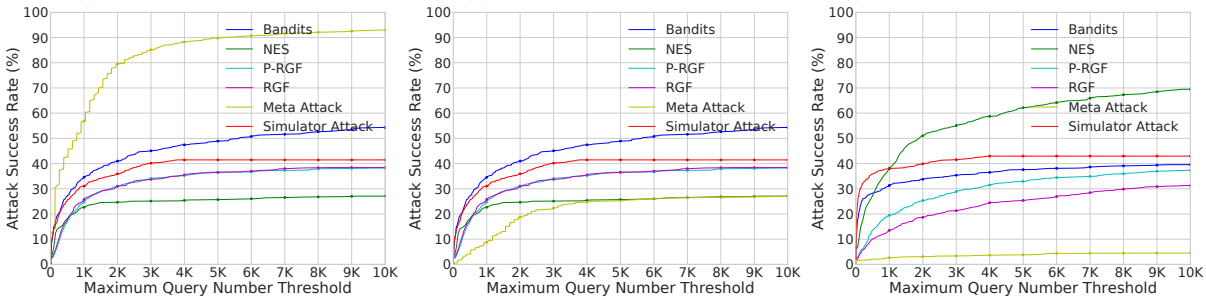
Figure 3: Comparisons of attack success rates at different limited maximum queries in TinyImageNet dataset.



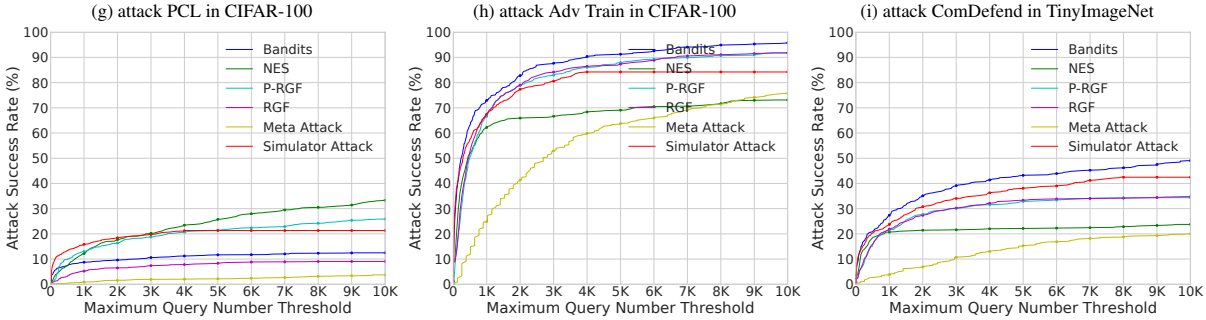
(a) attack ComDefend in CIFAR-10 (b) attack Feature Distillation in CIFAR-10 (c) attack PCL in CIFAR-10



(d) attack Adv Train in CIFAR-10 (e) attack ComDefend in CIFAR-100 (f) attack Feature Distillation in CIFAR-100

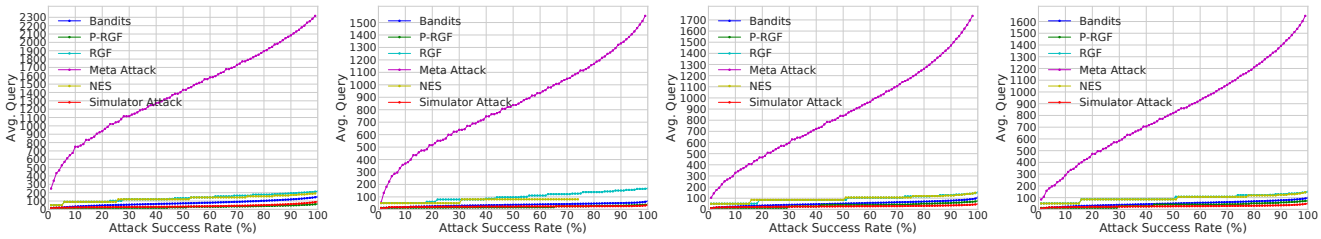


(g) attack PCL in CIFAR-100 (h) attack Adv Train in CIFAR-100 (i) attack ComDefend in TinyImageNet

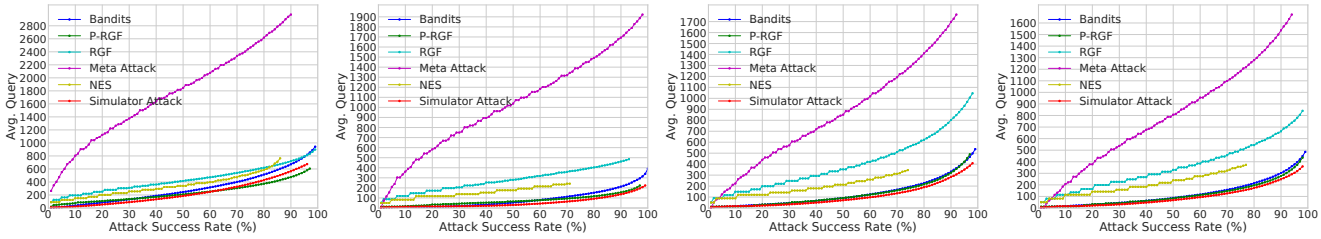


(j) attack Feature Distillation in TinyImageNet (k) attack PCL in TinyImageNet (l) attack Adv Train in TinyImageNet

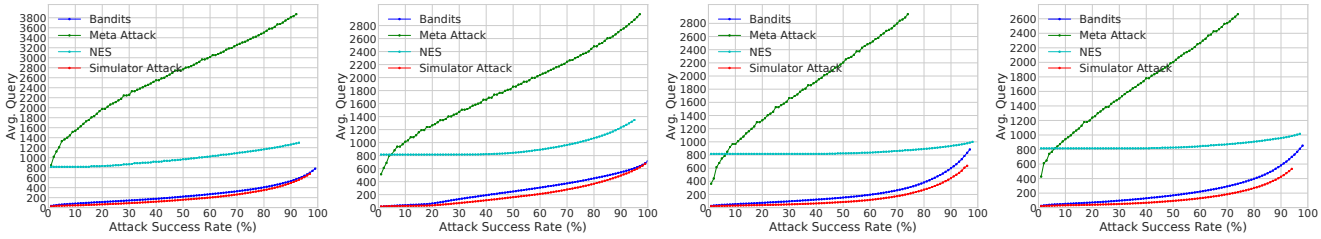
Figure 4: Comparisons of attack success rates at different maximum queries on defensive models with the ResNet-50 backbone. The experimental results are obtained by performing the untargeted attacks under ℓ_∞ norm.



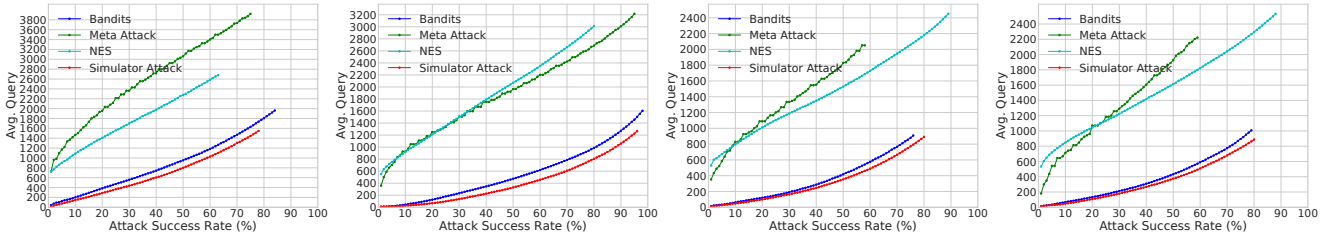
(a) untargeted ℓ_2 attack PyramidNet-272 (b) untargeted ℓ_2 attack GDAS (c) untargeted ℓ_2 attack WRN-28 (d) untargeted ℓ_2 attack WRN-40



(e) untargeted ℓ_∞ attack PyramidNet-272 (f) untargeted ℓ_∞ attack GDAS (g) untargeted ℓ_∞ attack WRN-28 (h) untargeted ℓ_∞ attack WRN-40

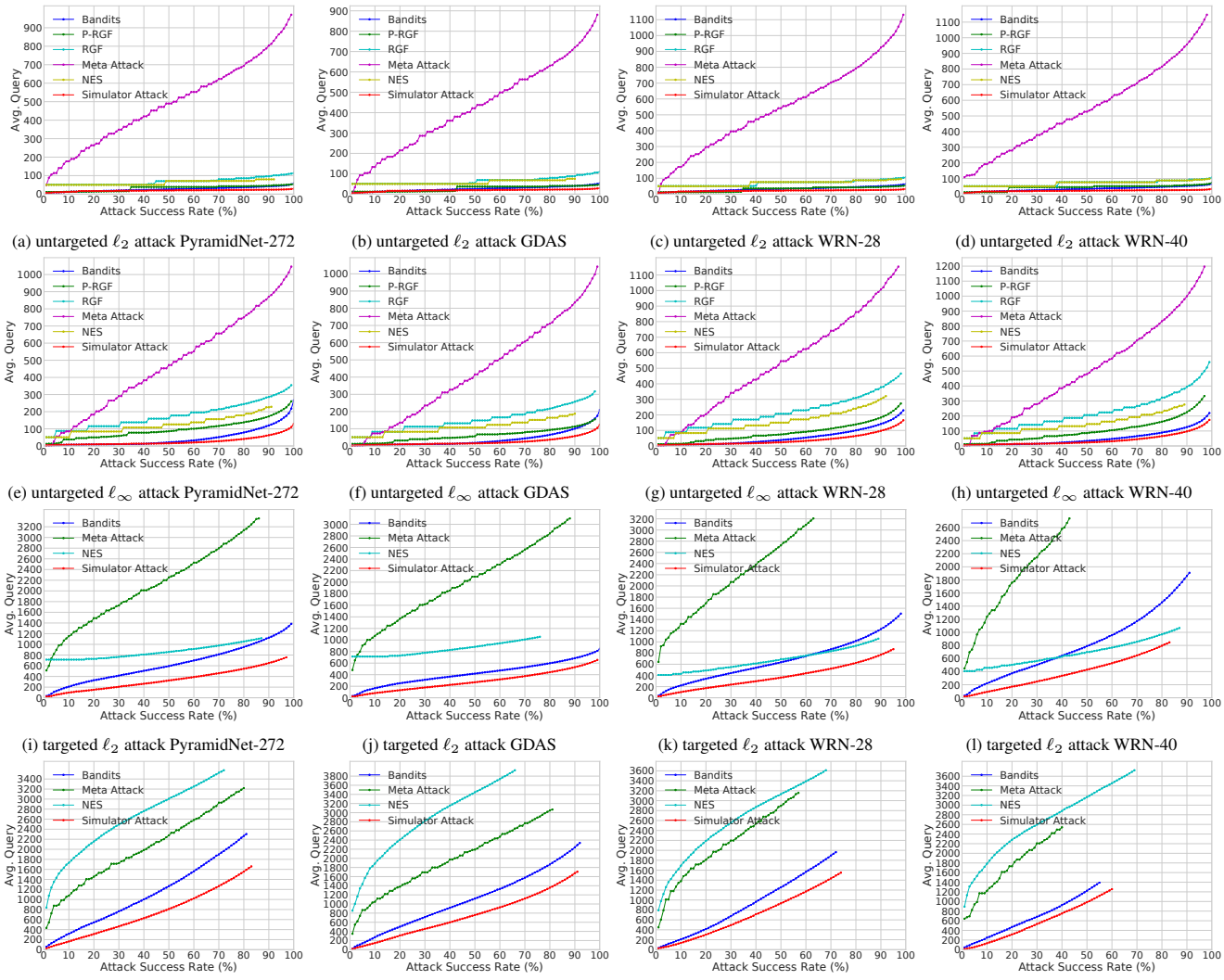


(i) targeted ℓ_2 attack PyramidNet-272 (j) targeted ℓ_2 attack GDAS (k) targeted ℓ_2 attack WRN-28 (l) targeted ℓ_2 attack WRN-40



(m) targeted ℓ_∞ attack PyramidNet-272 (n) targeted ℓ_∞ attack GDAS (o) targeted ℓ_∞ attack WRN-28 (p) targeted ℓ_∞ attack WRN-40

Figure 5: Comparisons of the average query per successful image at different desired success rates in CIFAR-10 dataset.



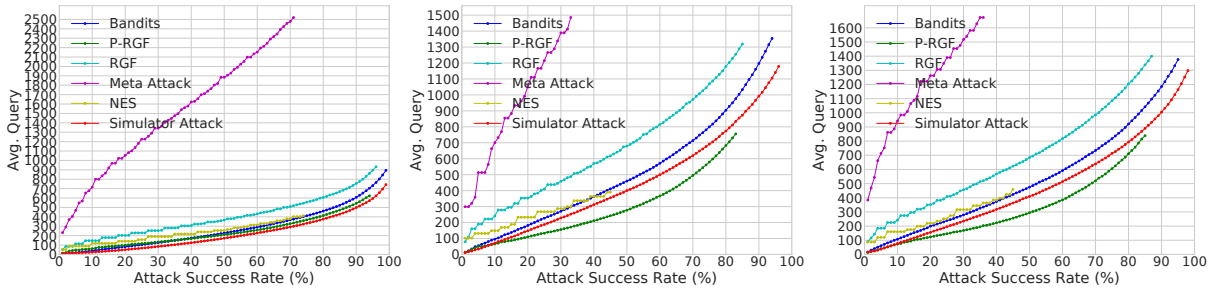
(a) untargeted ℓ_2 attack PyramidNet-272 (b) untargeted ℓ_2 attack GDAS (c) untargeted ℓ_2 attack WRN-28 (d) untargeted ℓ_2 attack WRN-40

(e) untargeted ℓ_∞ attack PyramidNet-272 (f) untargeted ℓ_∞ attack GDAS (g) untargeted ℓ_∞ attack WRN-28 (h) untargeted ℓ_∞ attack WRN-40

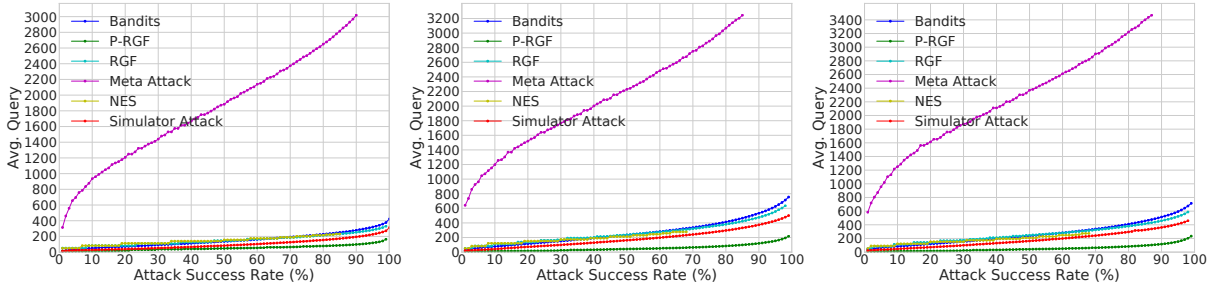
(i) targeted ℓ_2 attack PyramidNet-272 (j) targeted ℓ_2 attack GDAS (k) targeted ℓ_2 attack WRN-28 (l) targeted ℓ_2 attack WRN-40

(m) targeted ℓ_∞ attack PyramidNet-272 (n) targeted ℓ_∞ attack GDAS (o) targeted ℓ_∞ attack WRN-28 (p) targeted ℓ_∞ attack WRN-40

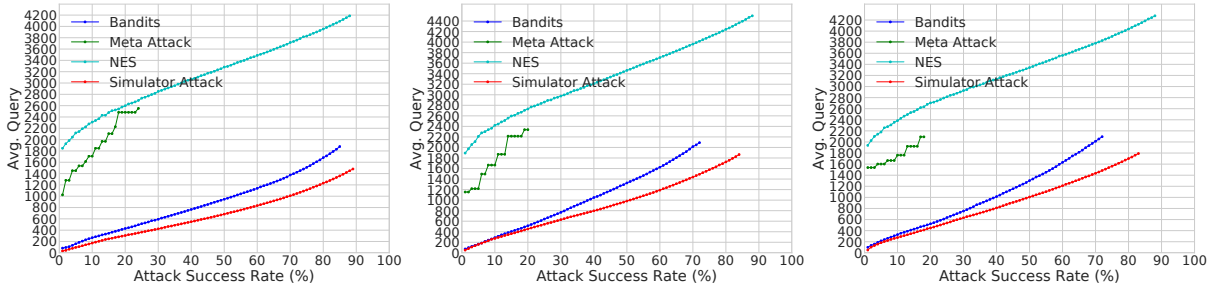
Figure 6: Comparisons of the average query per successful image at different desired success rates in CIFAR-100 dataset.



(a) untargeted ℓ_∞ attack DenseNet-121 (b) untargeted ℓ_∞ attack ResNeXt-101(32×4d) (c) untargeted ℓ_∞ attack ResNeXt-101(64×4d)

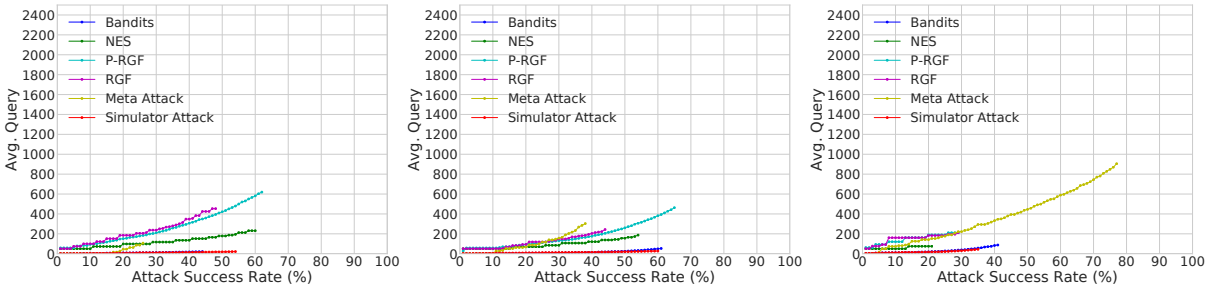


(d) untargeted ℓ_2 attack DenseNet-121 (e) untargeted ℓ_2 attack ResNeXt-101(32×4d) (f) untargeted ℓ_2 attack ResNeXt-101(64×4d)

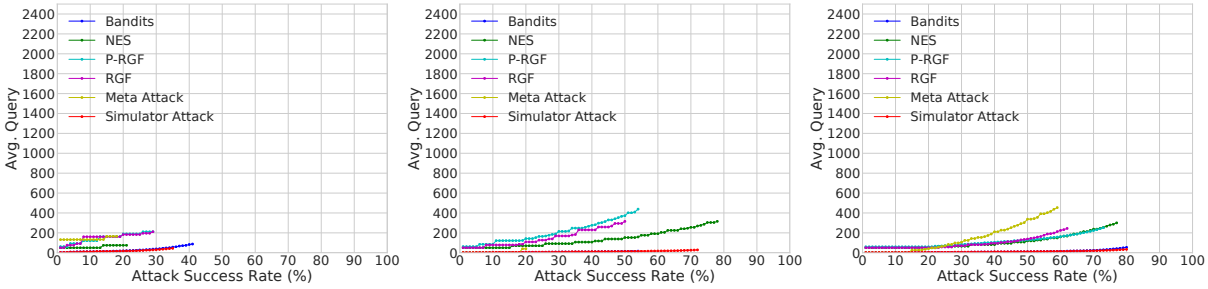


(g) targeted ℓ_2 attack DenseNet-121 (h) targeted ℓ_2 attack ResNeXt-101(32×4d) (i) targeted ℓ_2 attack ResNeXt-101(64×4d)

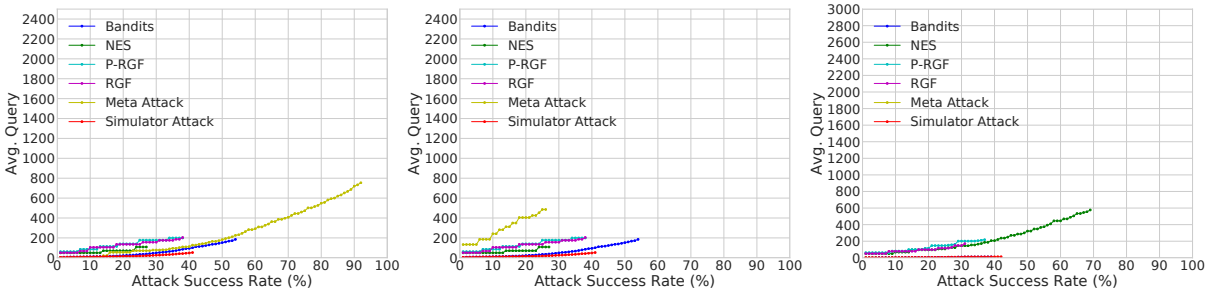
Figure 7: Comparisons of the average query per successful image at different desired success rates in TinyImageNet dataset.



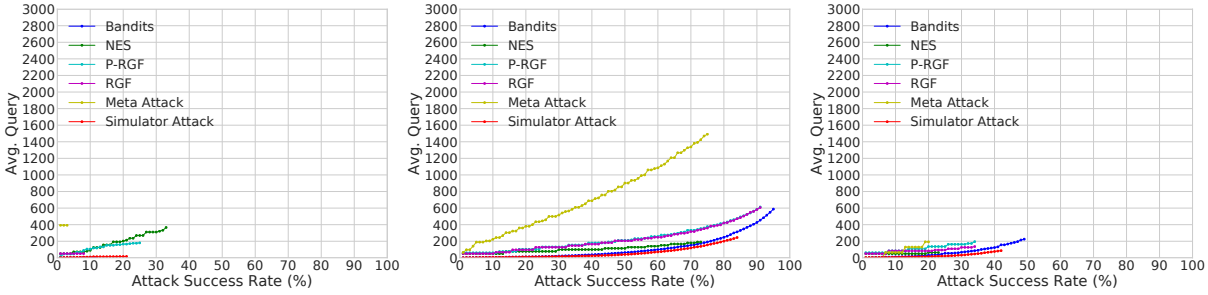
(a) attack ComDefend in CIFAR-10 (b) attack Feature Distillation in CIFAR-10 (c) attack PCL in CIFAR-10



(d) attack Adv Train in CIFAR-10 (e) attack ComDefend in CIFAR-100 (f) attack Feature Distillation in CIFAR-100

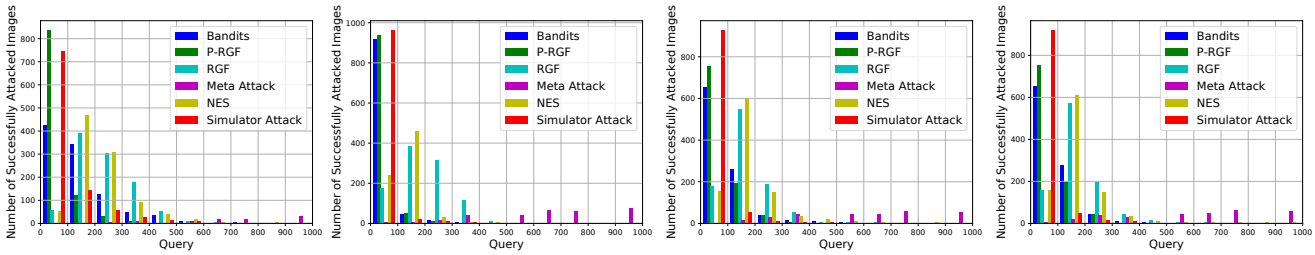


(g) attack PCL in CIFAR-100 (h) attack Adv Train in CIFAR-100 (i) attack ComDefend in TinyImageNet

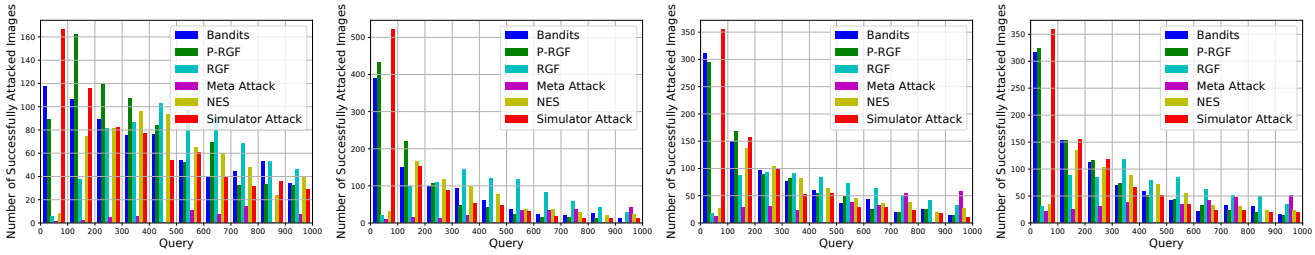


(j) attack Feature Distillation in TinyImageNet (k) attack PCL in TinyImageNet (l) attack Adv Train in TinyImageNet

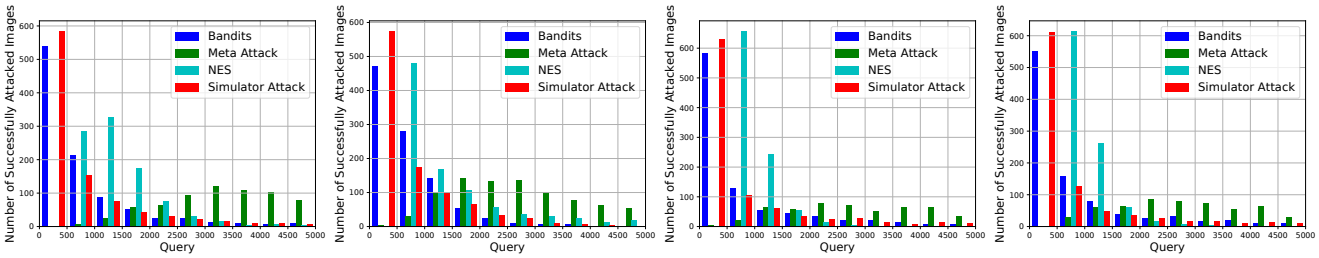
Figure 8: Comparisons of the average query per successful image at different desired success rates on defensive models with the backbone of ResNet-50. The experimental results are obtained by performing the untargeted attacks under ℓ_∞ norm.



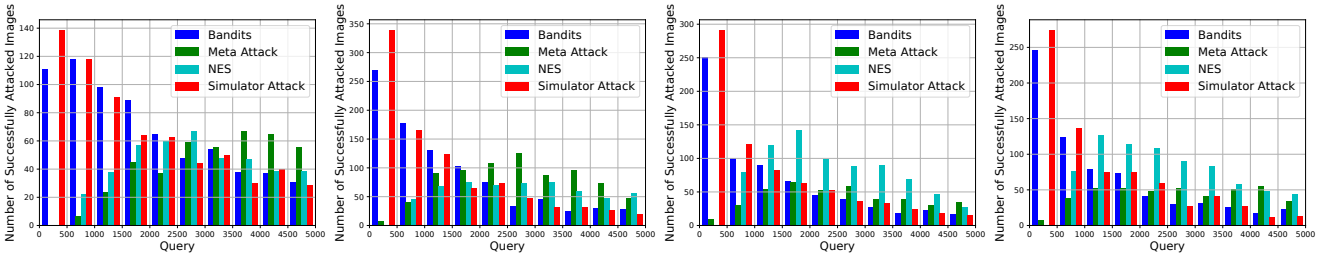
(a) untargeted ℓ_2 attack PyramidNet-272 (b) untargeted ℓ_2 attack GDAS (c) untargeted ℓ_2 attack WRN-28 (d) untargeted ℓ_2 attack WRN-40



(e) untargeted ℓ_∞ attack PyramidNet-272 (f) untargeted ℓ_∞ attack GDAS (g) untargeted ℓ_∞ attack WRN-28 (h) untargeted ℓ_∞ attack WRN-40

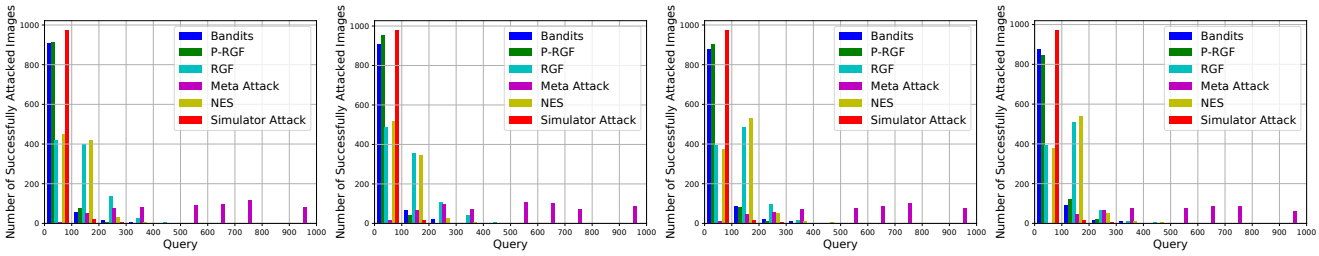


(i) targeted ℓ_2 attack PyramidNet-272 (j) targeted ℓ_2 attack GDAS (k) targeted ℓ_2 attack WRN-28 (l) targeted ℓ_2 attack WRN-40

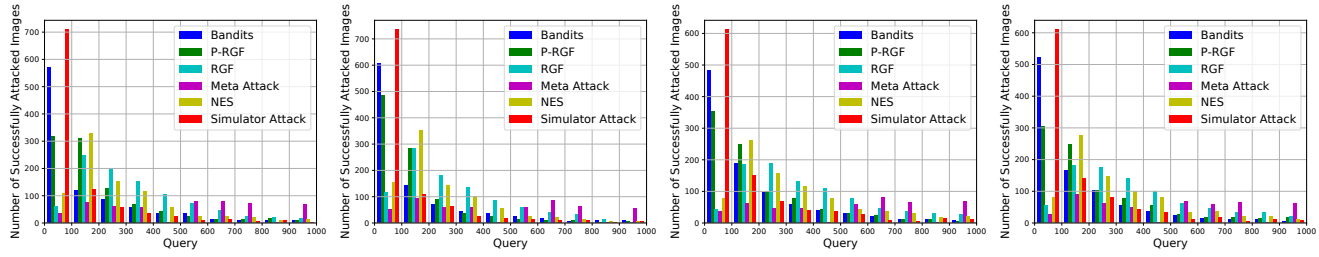


(m) targeted ℓ_∞ attack PyramidNet-272 (n) targeted ℓ_∞ attack GDAS (o) targeted ℓ_∞ attack WRN-28 (p) targeted ℓ_∞ attack WRN-40

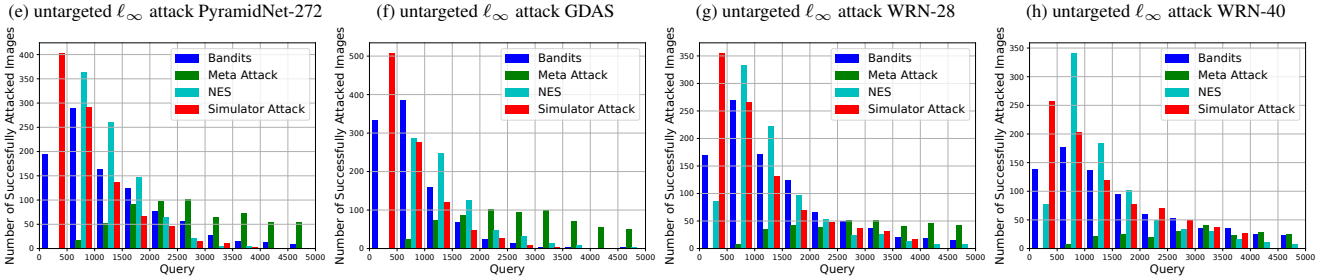
Figure 9: The histogram of query number in the CIFAR-10 dataset.



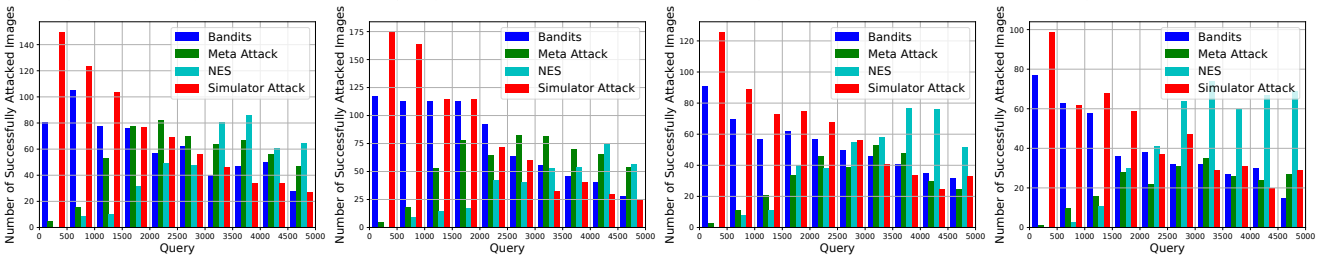
(a) untargeted ℓ_2 attack PyramidNet-272 (b) untargeted ℓ_2 attack GDAS (c) untargeted ℓ_2 attack WRN-28 (d) untargeted ℓ_2 attack WRN-40



(e) untargeted ℓ_∞ attack PyramidNet-272 (f) untargeted ℓ_∞ attack GDAS (g) untargeted ℓ_∞ attack WRN-28 (h) untargeted ℓ_∞ attack WRN-40

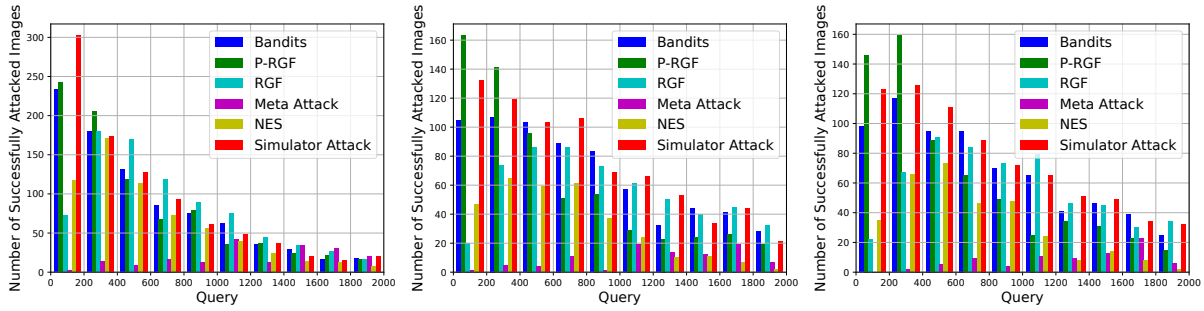


(i) targeted ℓ_2 attack PyramidNet-272 (j) targeted ℓ_2 attack GDAS (k) targeted ℓ_2 attack WRN-28 (l) targeted ℓ_2 attack WRN-40

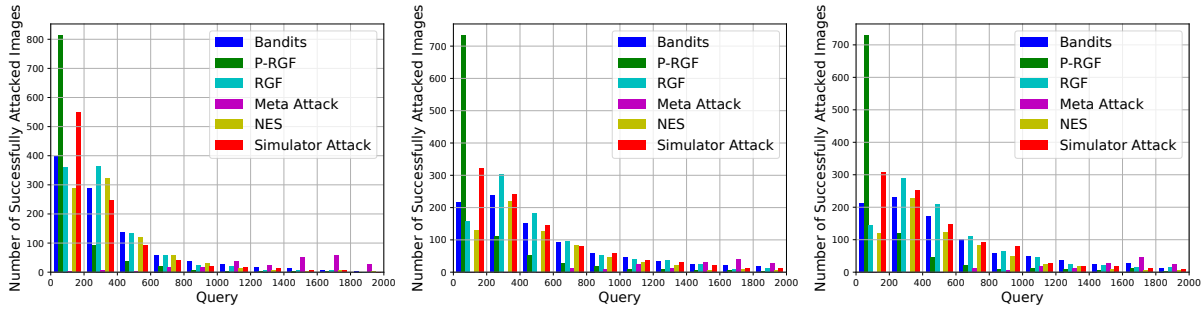


(m) targeted ℓ_∞ attack PyramidNet-272 (n) targeted ℓ_∞ attack GDAS (o) targeted ℓ_∞ attack WRN-28 (p) targeted ℓ_∞ attack WRN-40

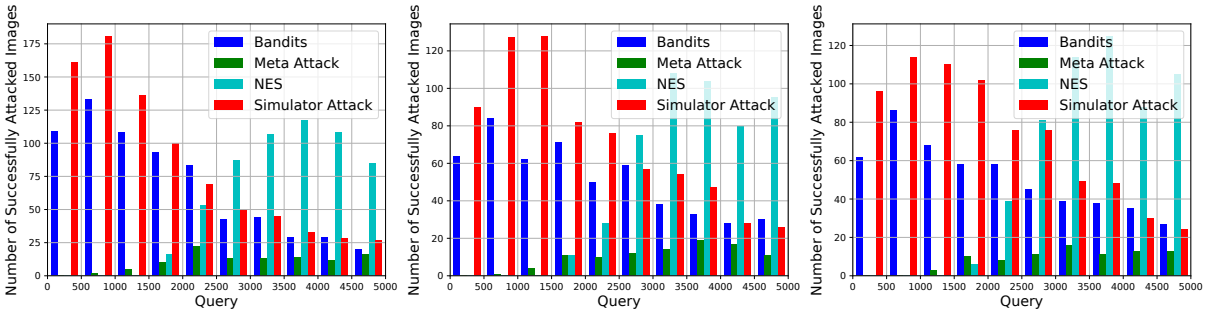
Figure 10: The histogram of query number in the CIFAR-100 dataset.



(a) untargeted ℓ_∞ attack DenseNet-121 (b) untargeted ℓ_∞ attack ResNext-101(32×4d) (c) untargeted ℓ_∞ attack ResNext-101(64×4d)

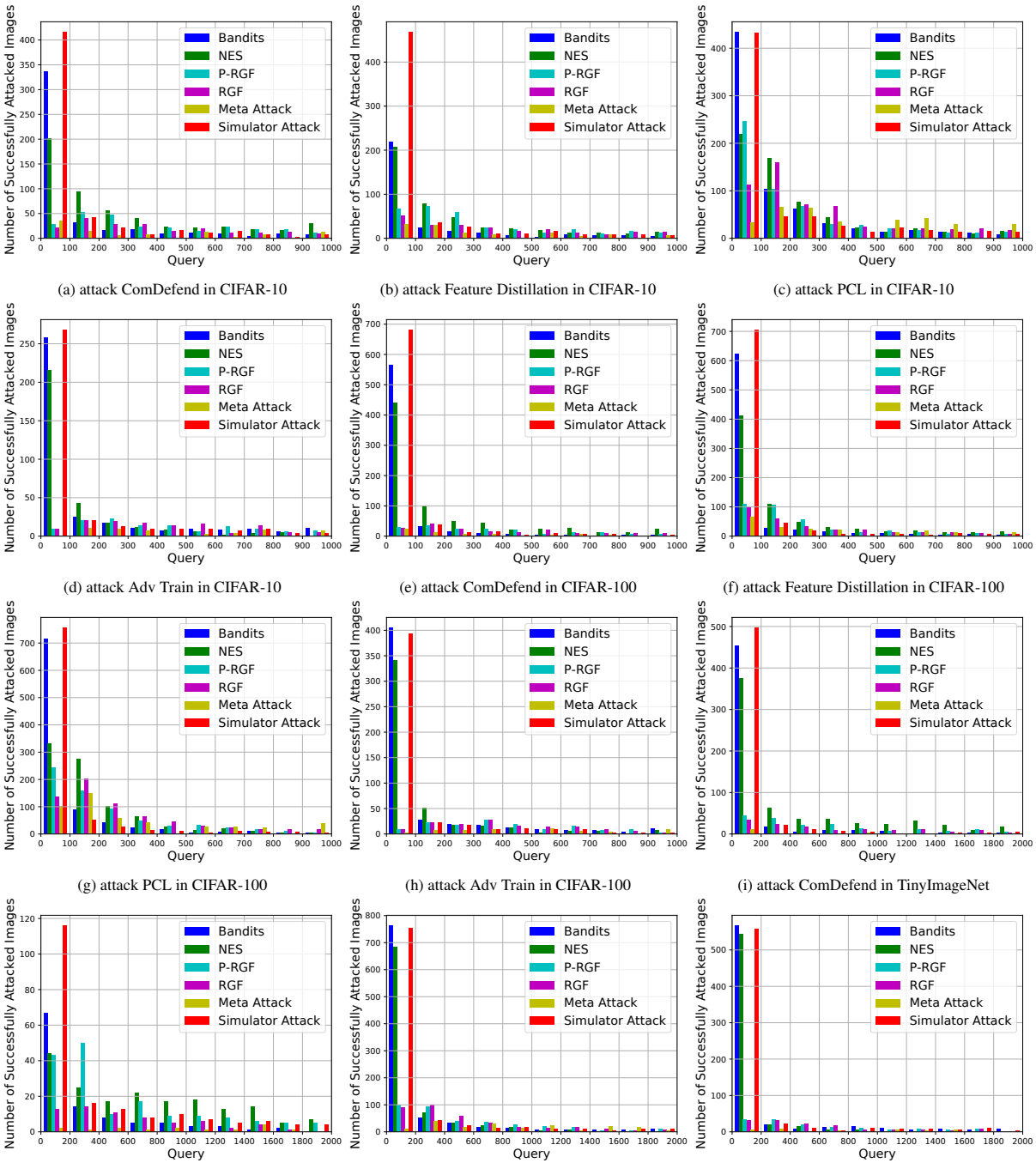


(d) untargeted ℓ_2 attack DenseNet-121 (e) untargeted ℓ_2 attack ResNext-101(32×4d) (f) untargeted ℓ_2 attack ResNext-101(64×4d)



(g) targeted ℓ_2 attack DenseNet-121 (h) targeted ℓ_2 attack ResNext-101(32×4d) (i) targeted ℓ_2 attack ResNext-101(64×4d)

Figure 11: The histogram of query number in the TinyImageNet dataset.



(a) attack ComDefend in CIFAR-10 (b) attack Feature Distillation in CIFAR-10 (c) attack PCL in CIFAR-10

(d) attack Adv Train in CIFAR-10 (e) attack ComDefend in CIFAR-100 (f) attack Feature Distillation in CIFAR-100

(g) attack PCL in CIFAR-100 (h) attack Adv Train in CIFAR-100 (i) attack ComDefend in TinyImageNet

(j) attack Feature Distillation in TinyImageNet (k) attack PCL in TinyImageNet (l) attack Adv Train in TinyImageNet

Figure 12: The histogram of query number on defensive models with the backbone of ResNet-50. The experimental results are obtained by performing the untargeted attacks under ℓ_∞ norm.

References

[1] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[2] Jiawei Du, Hu Zhang, Joey Tianyi Zhou, Yi Yang, and Jiashi Feng. Query-efficient meta attack to deep neural networks. In *International Conference on Learning Representations*, 2020. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[4] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2137–2146. PMLR, 10–15 Jul 2018. 1, 2

[5] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *International Conference on Learning Representations*, 2019. 1

[6] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. Comdefend: An efficient image compression model to defend adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6084–6092, 2019. 2

[7] P. Diederik Kingma and Lei Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 3

[8] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868. IEEE, 2019. 2

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2

[10] Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Adversarial defense by restricting the hidden space of deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3385–3394, 2019. 2

[11] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. 1