ROME is Forged in Adversity: RObust Distilled Datasets via InforMation BottlenEck

Zheng Zhou[1], Wenquan Feng[1], Qiaosheng Zhang[2 3], Shuchang Lyu[1 *], Qi Zhao[1], Guangliang Cheng[4]

[1]Beihang University, [2]Shanghai Artificial Intelligence Laboratory, [3]Shanghai Innovation Institute, [4]University of Liverpool

*Corresponding Author

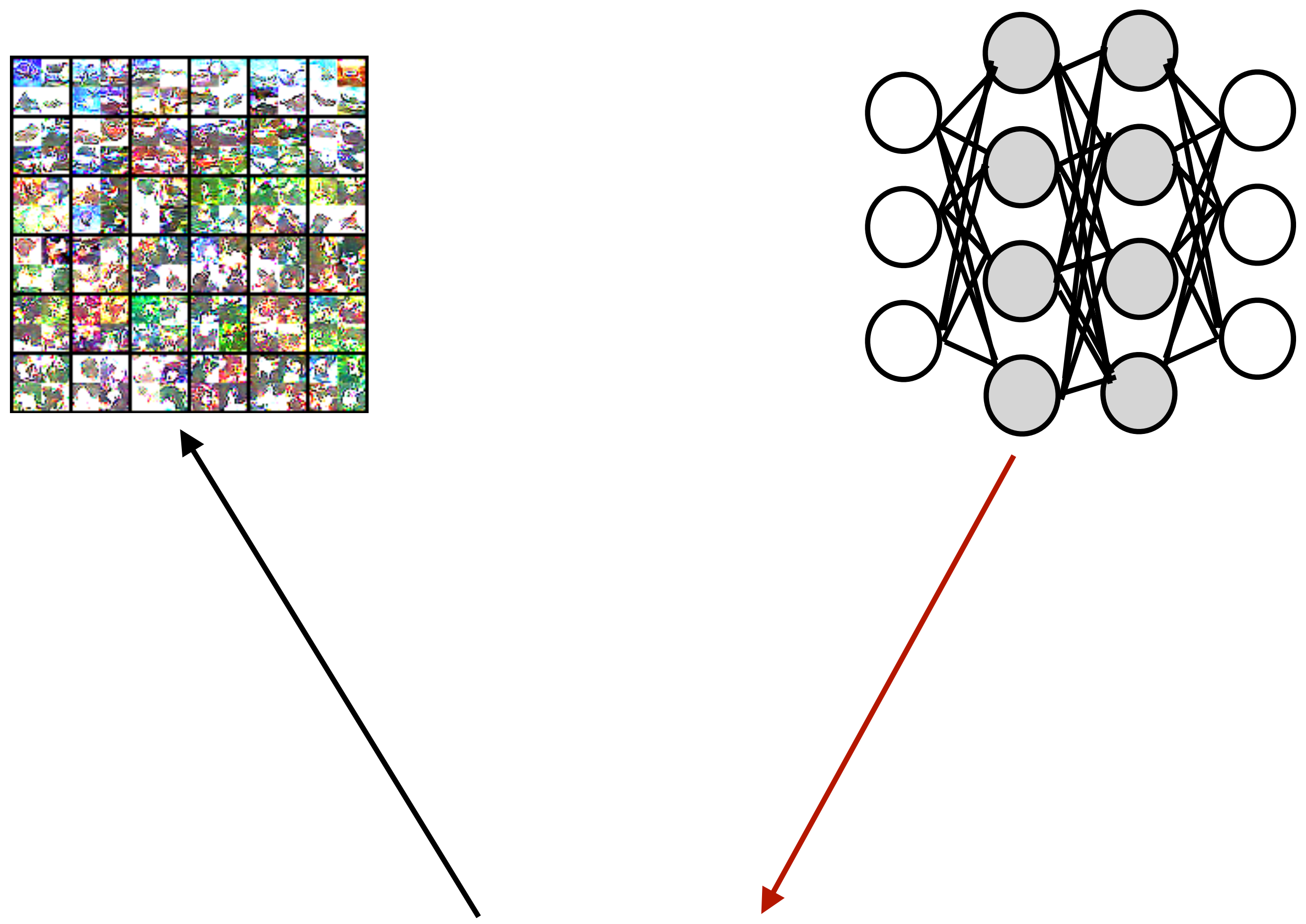
Code Project Page Contact us

What is Dataset Distillation?

Dataset distillation compresses large datasets into compact synthetic subsets, significantly reducing training time and computation while maintaining model performance.

What is Dataset Distillation?

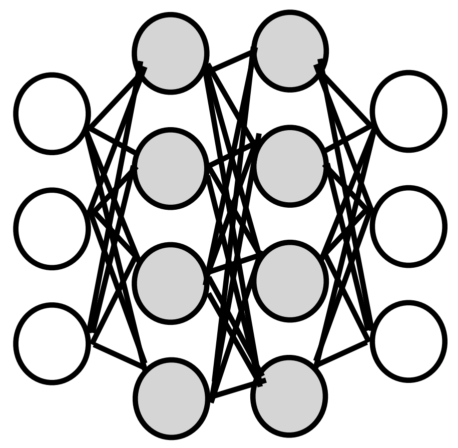
Dataset distillation compresses large datasets into compact synthetic subsets, significantly reducing training time and computation while maintaining model performance.



Train on distilled dataset

What is Dataset Distillation?

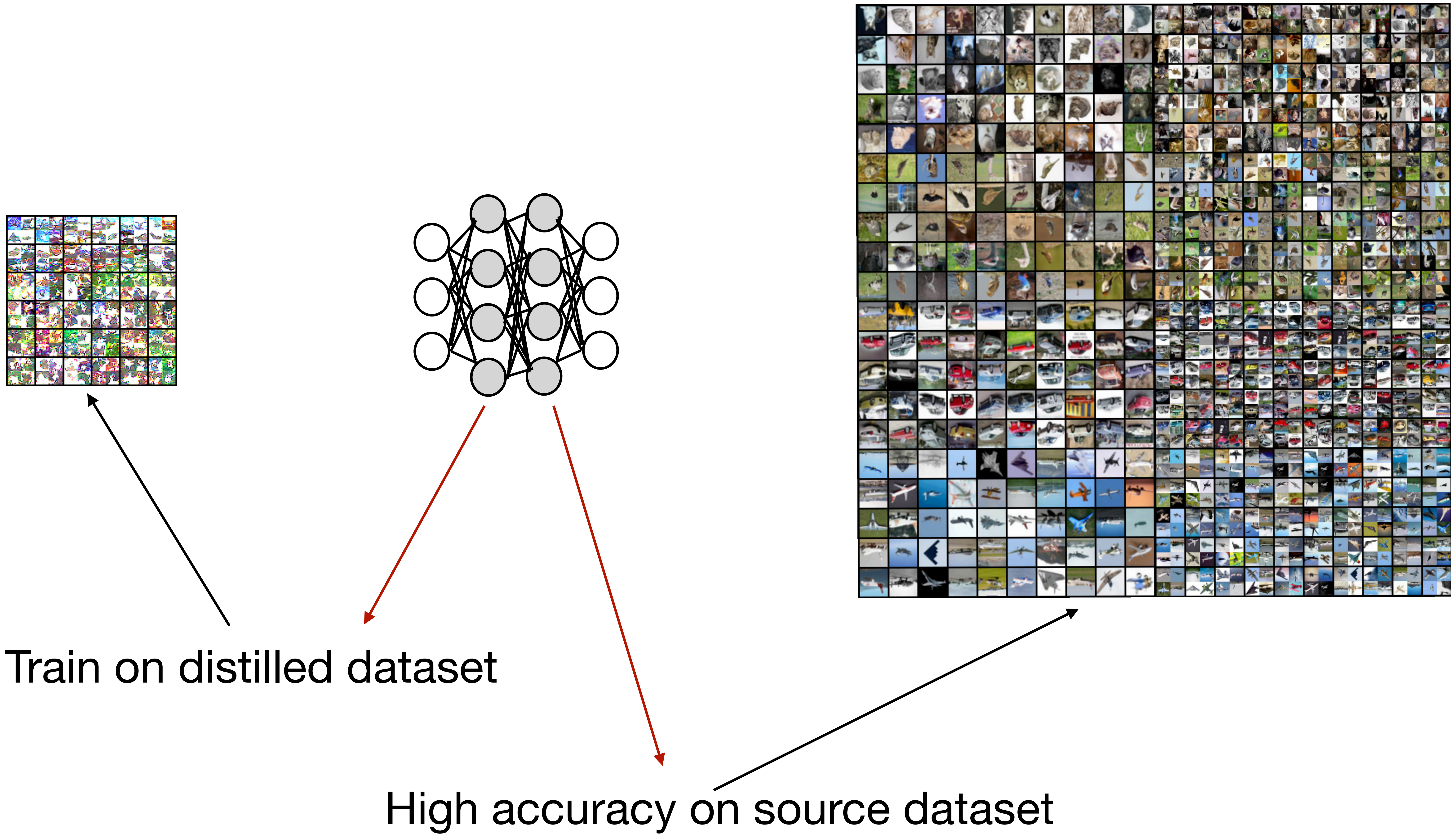
Dataset distillation compresses large datasets into compact synthetic subsets, significantly reducing training time and computation while maintaining model performance.



High accuracy on source dataset

What is Dataset Distillation?

Dataset distillation compresses large datasets into compact synthetic subsets, significantly reducing training time and computation while maintaining model performance.

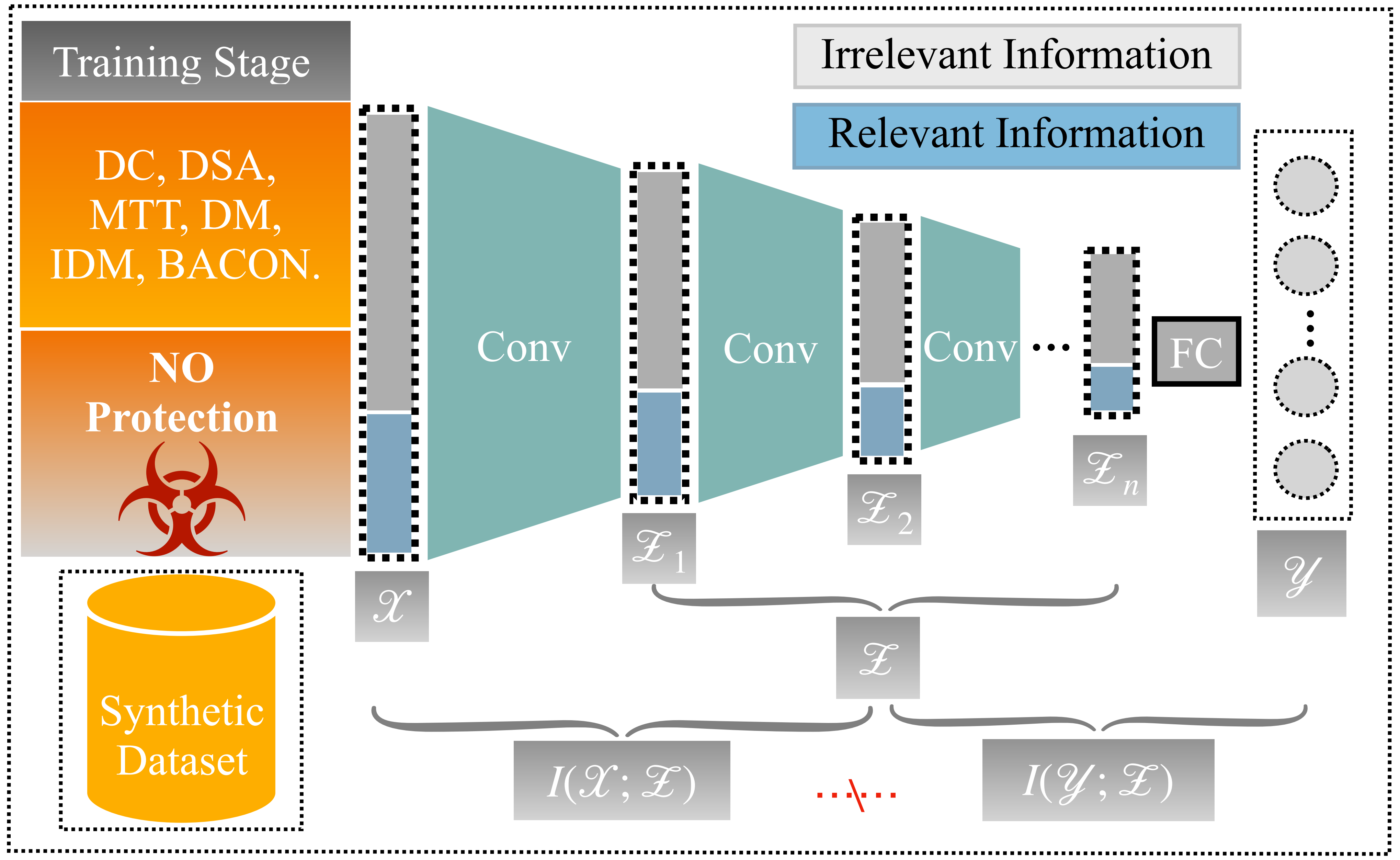


Efficiency Without Security

Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.

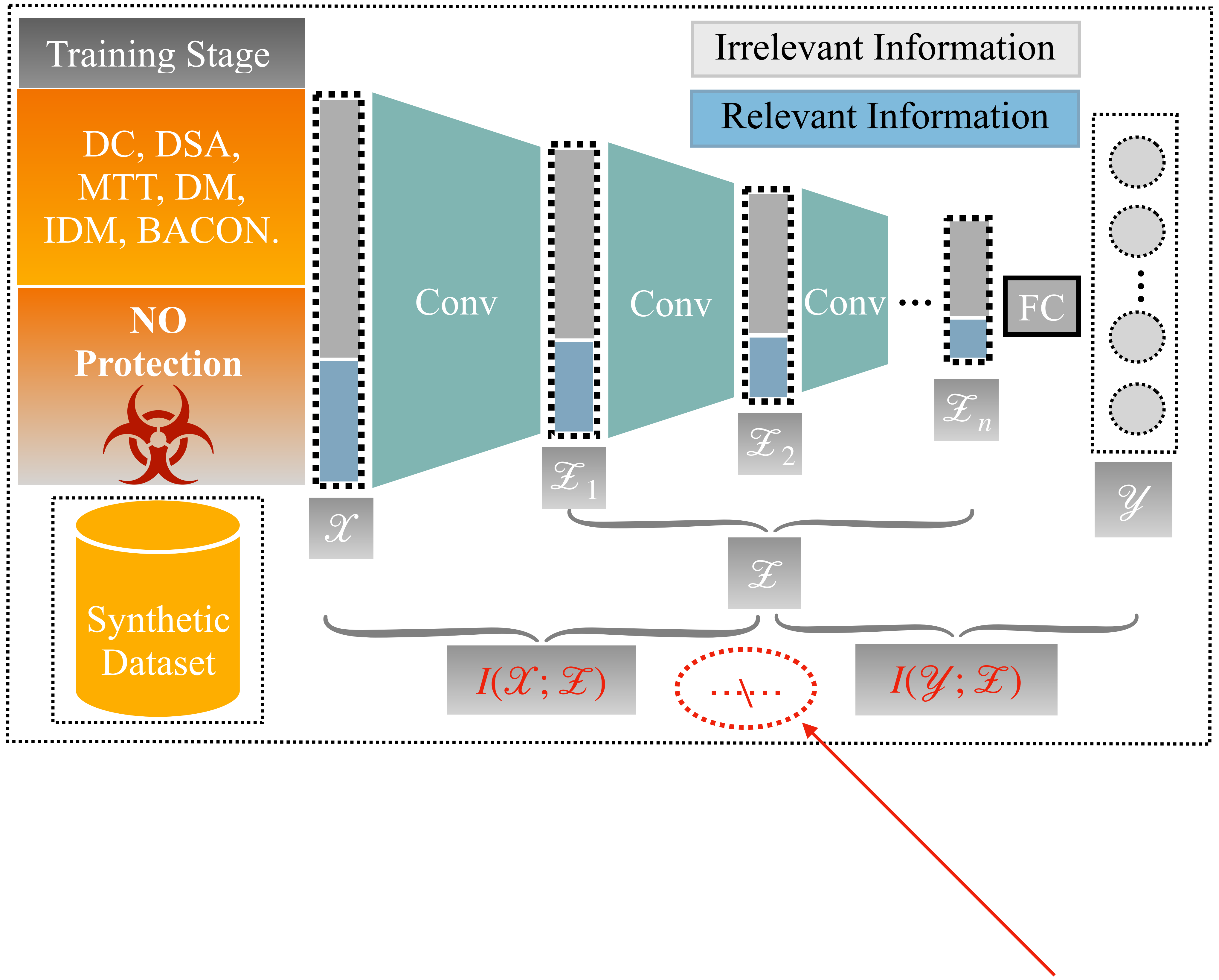
Efficiency Without Security

Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.



Efficiency Without Security

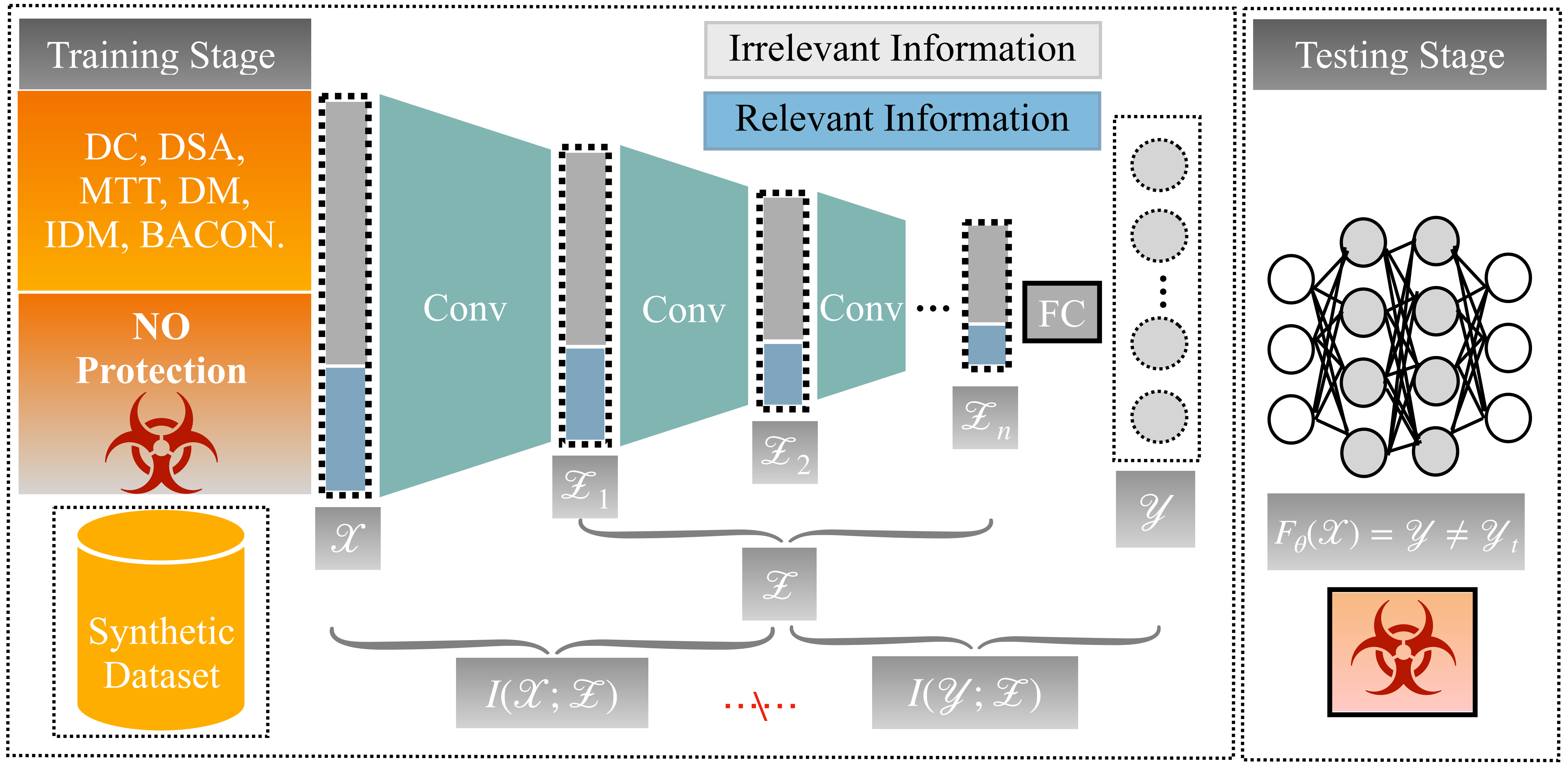
Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.



No **mutual information** is modeled among $\mathscr{X}$, $\mathscr{Z}$ and $\mathscr{Y}$

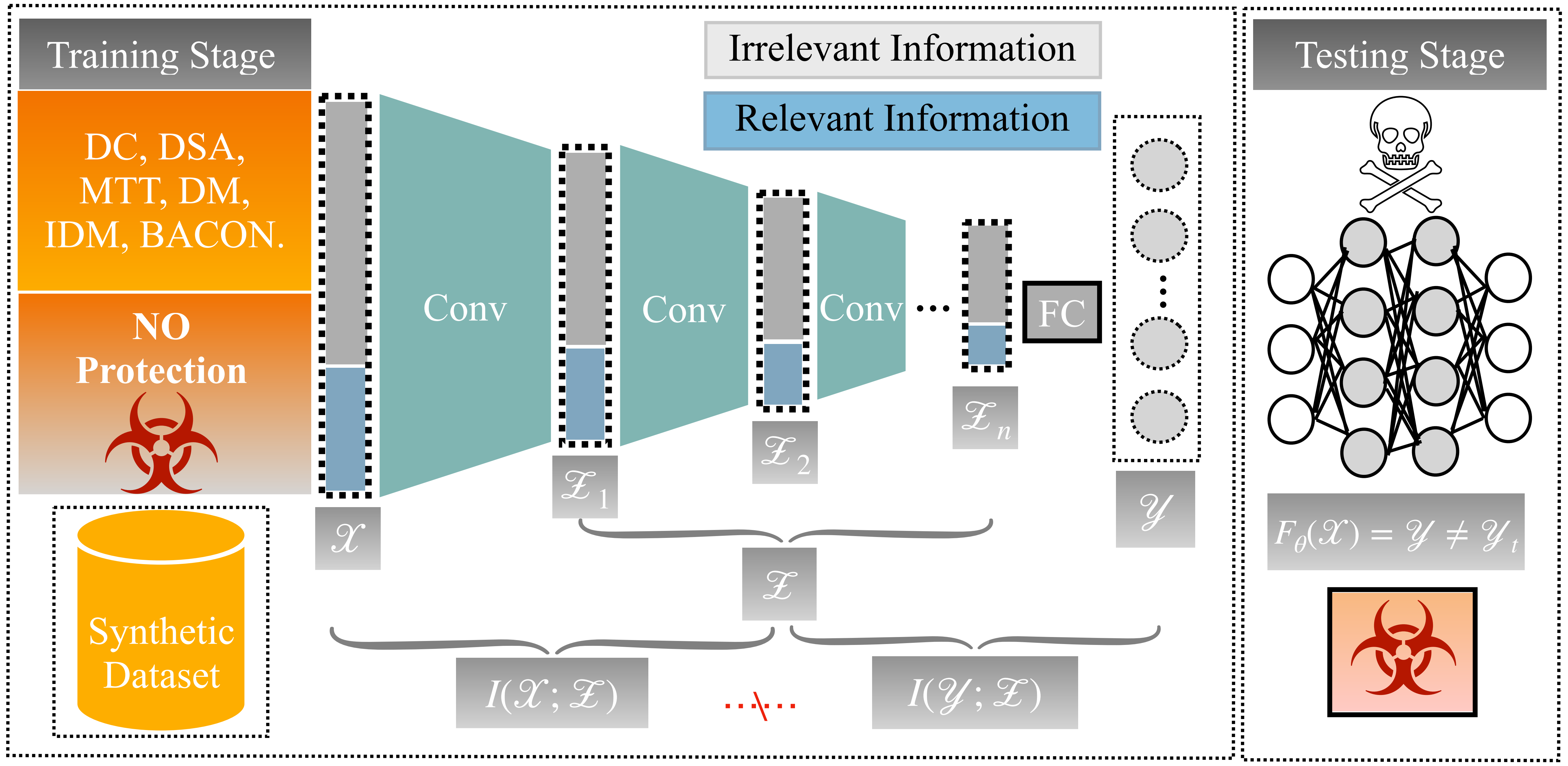
Efficiency Without Security

Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.



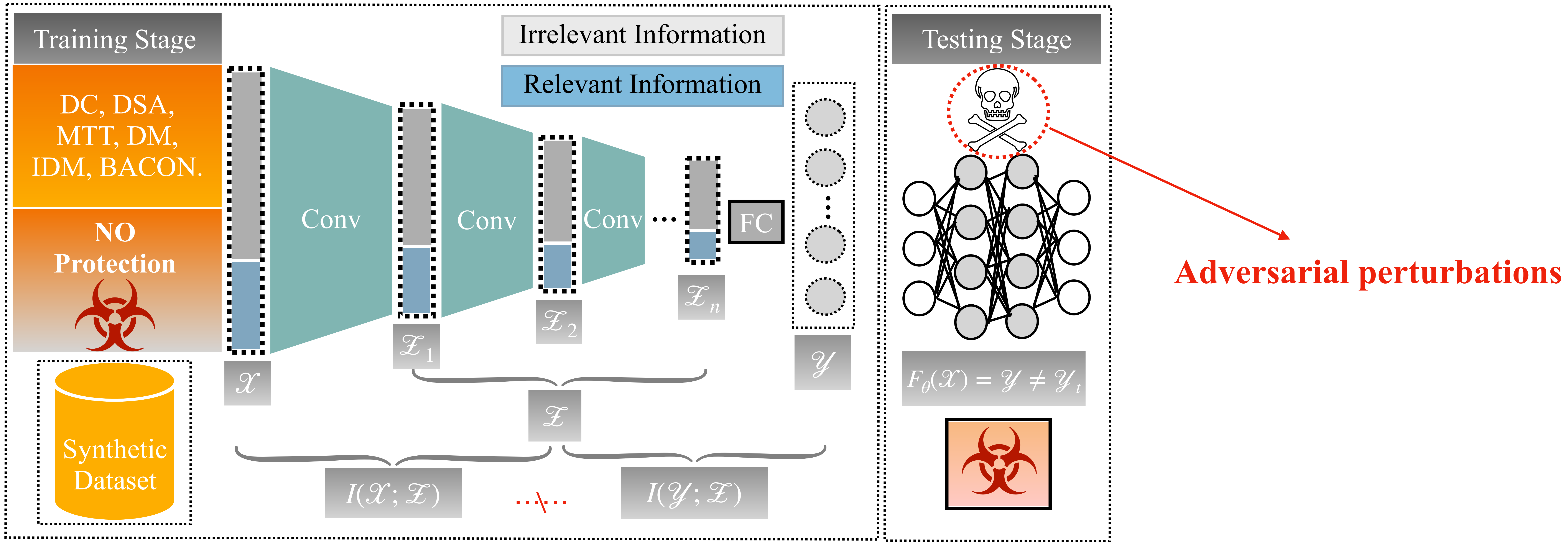
Efficiency Without Security

Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.



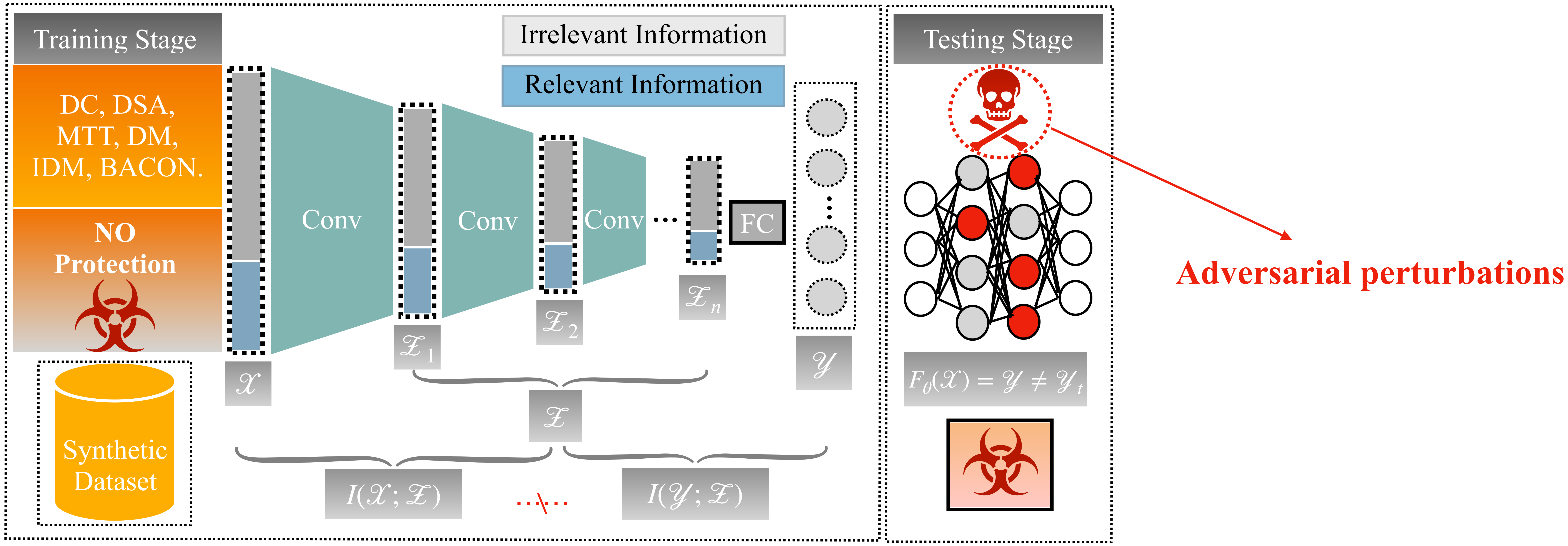
Efficiency Without Security

Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.



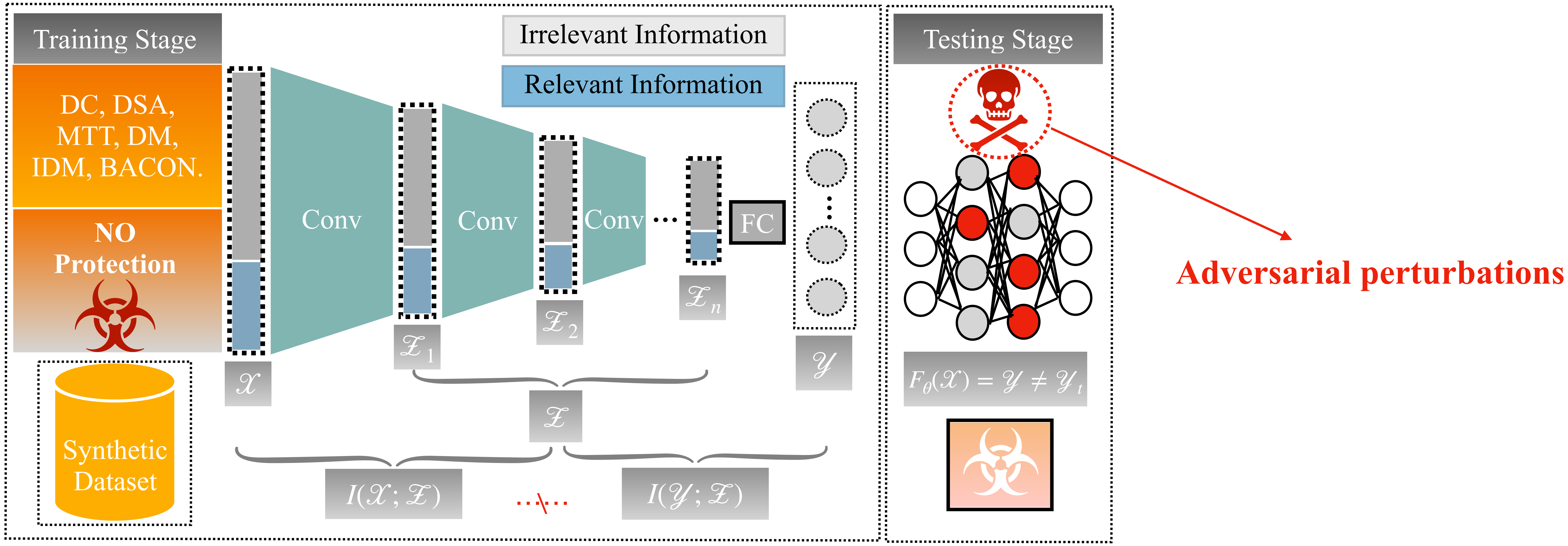
Efficiency Without Security

Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.



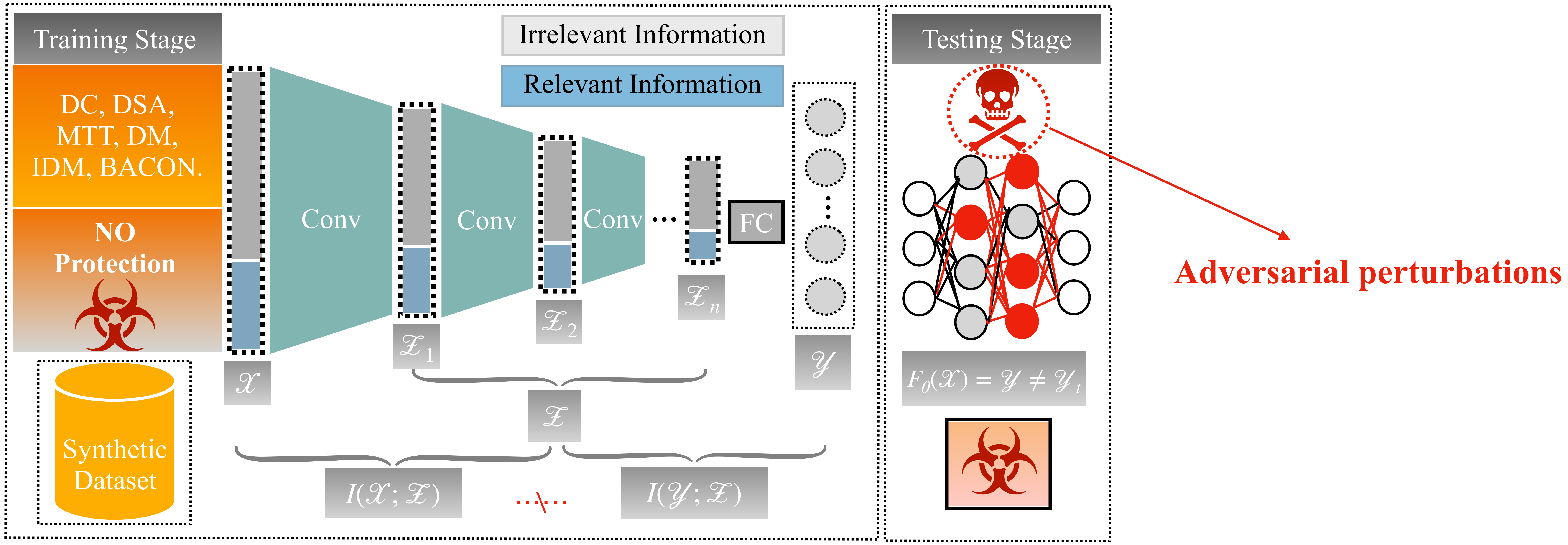
Efficiency Without Security

Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.



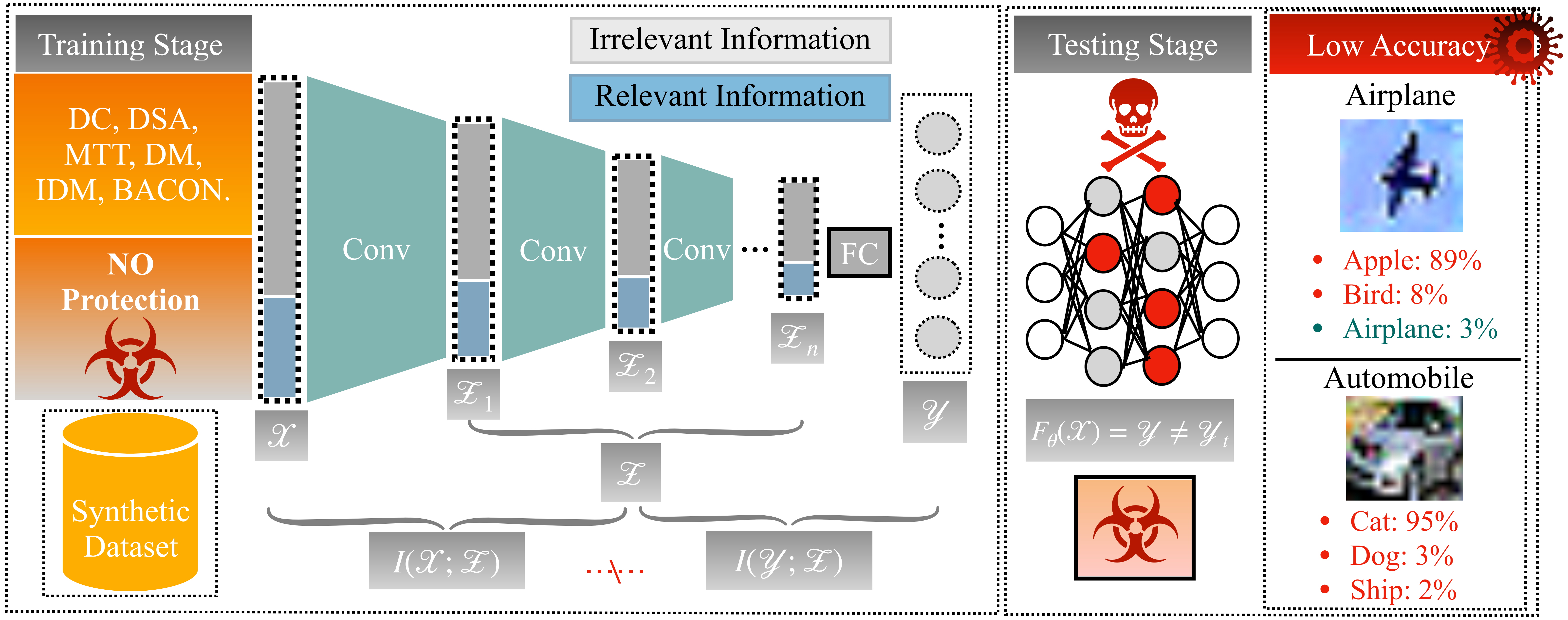
Efficiency Without Security

Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.



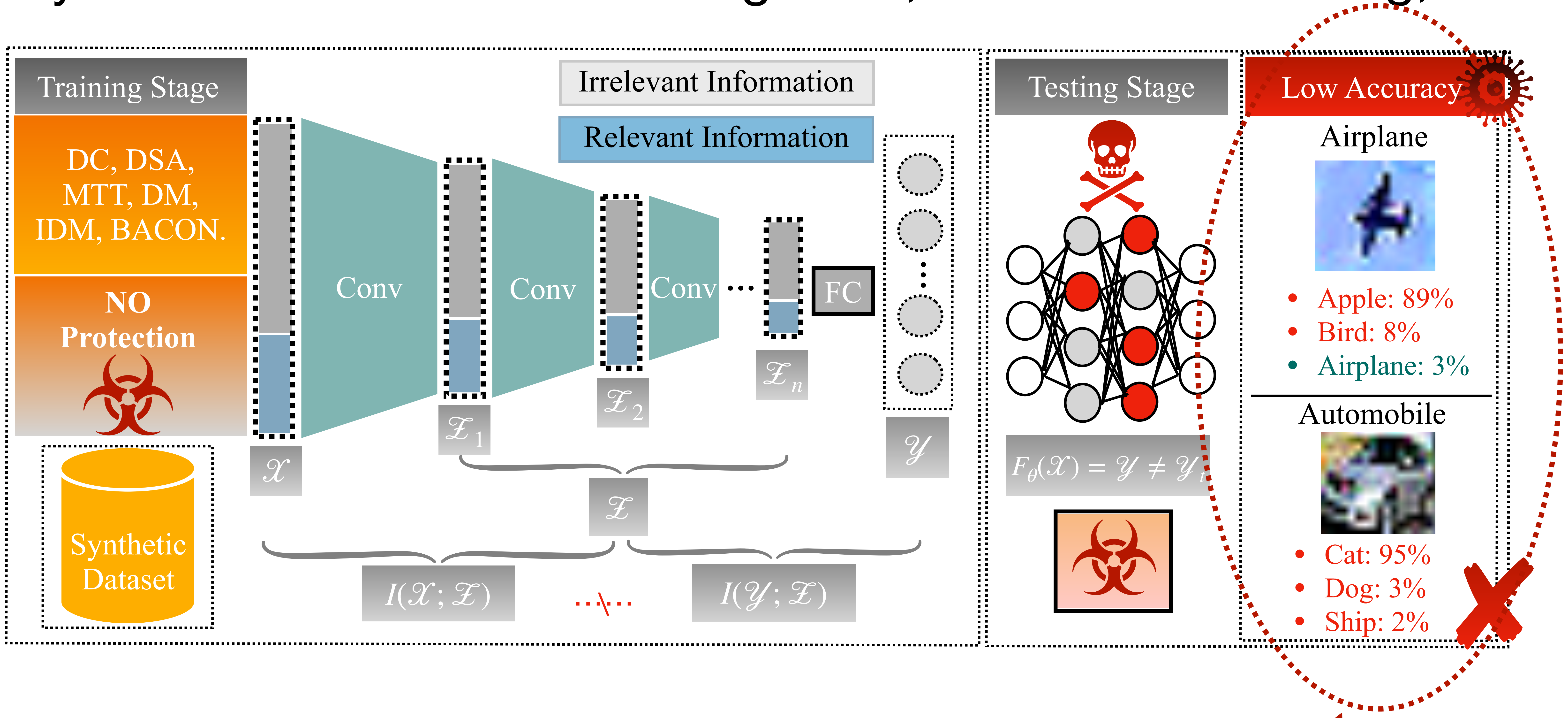
Efficiency Without Security

Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.



Efficiency Without Security

Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.



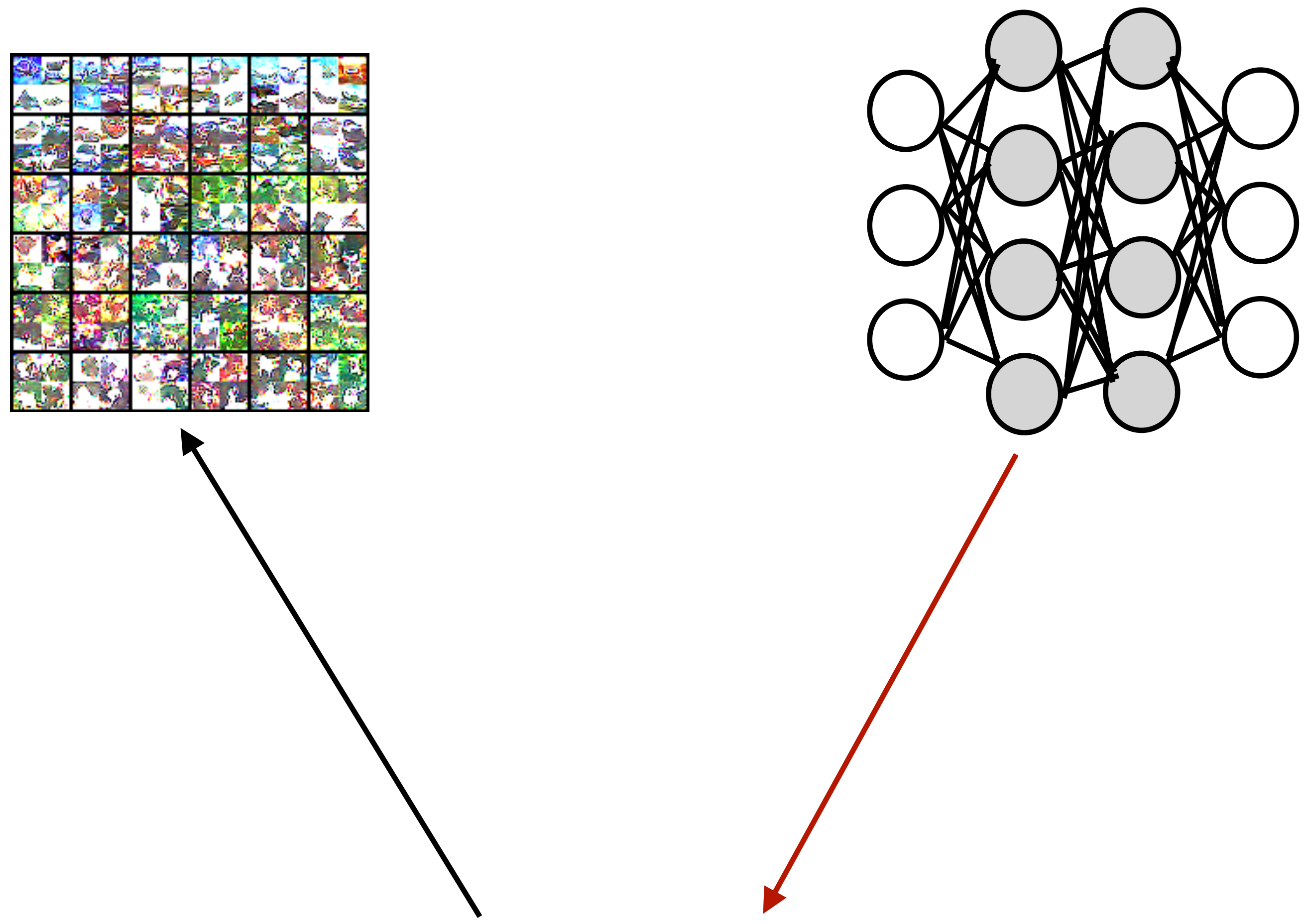
Dataset distillation improves efficiency, but not **robustness**.

How to enhance the robustness of models?

Adversarial robustness is a key research focus. A common way to improve it is adversarial training, but this method is costly and hard to apply in data-efficient settings like dataset distillation.

How to enhance the robustness of models?

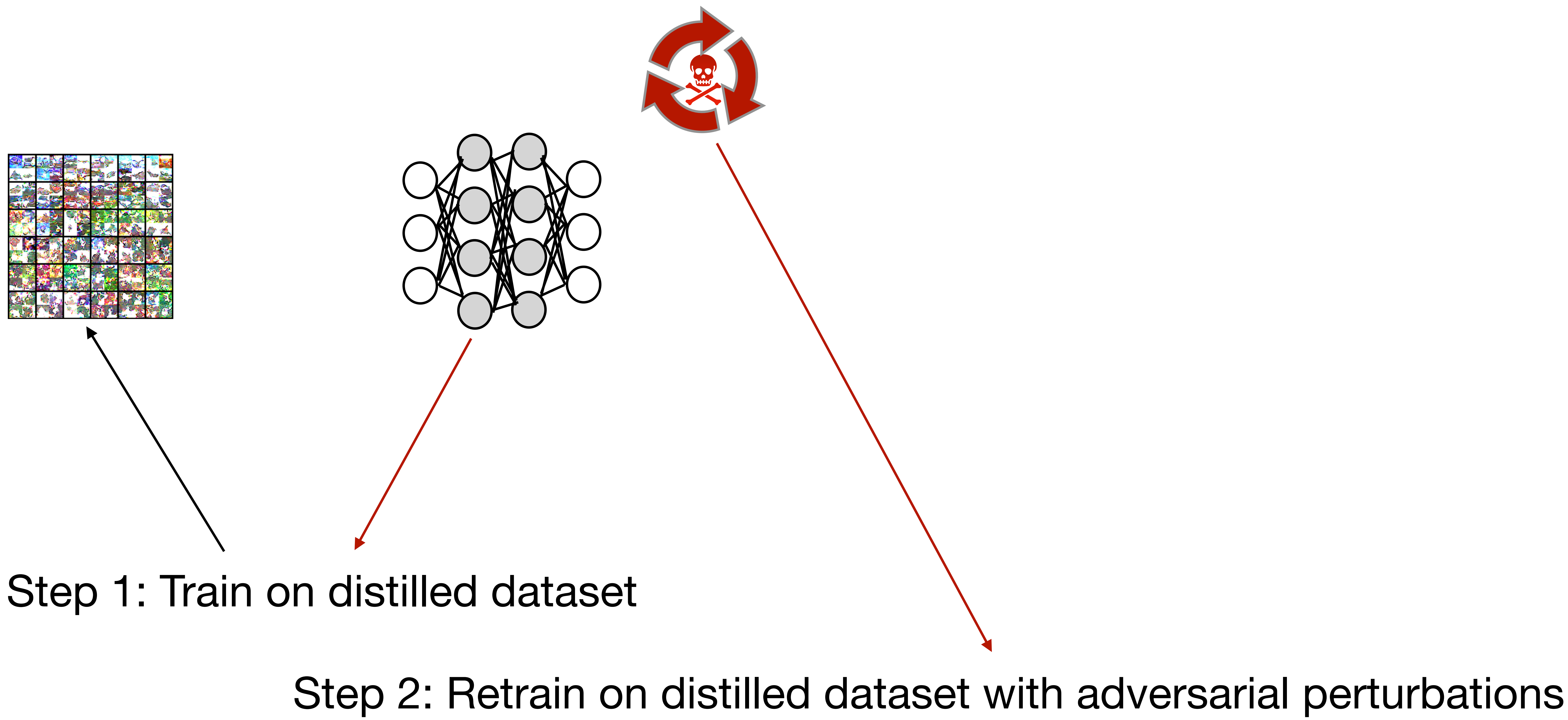
Adversarial robustness is a key research focus. A common way to improve it is adversarial training, but this method is costly and hard to apply in data-efficient settings like dataset distillation.



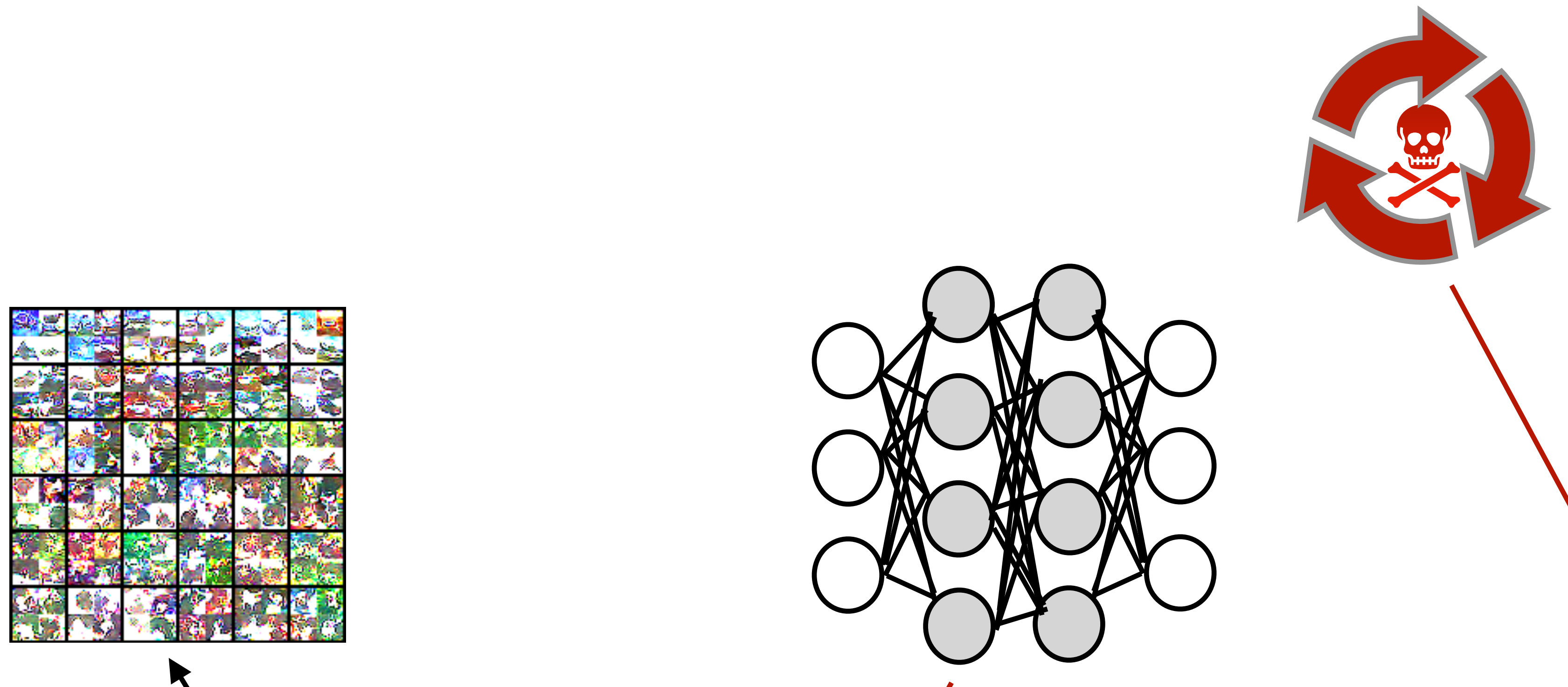
Step 1: Train on distilled dataset

How to enhance the robustness of models?

Adversarial robustness is a key research focus. A common way to improve it is adversarial training, but this method is costly and hard to apply in data-efficient settings like dataset distillation.



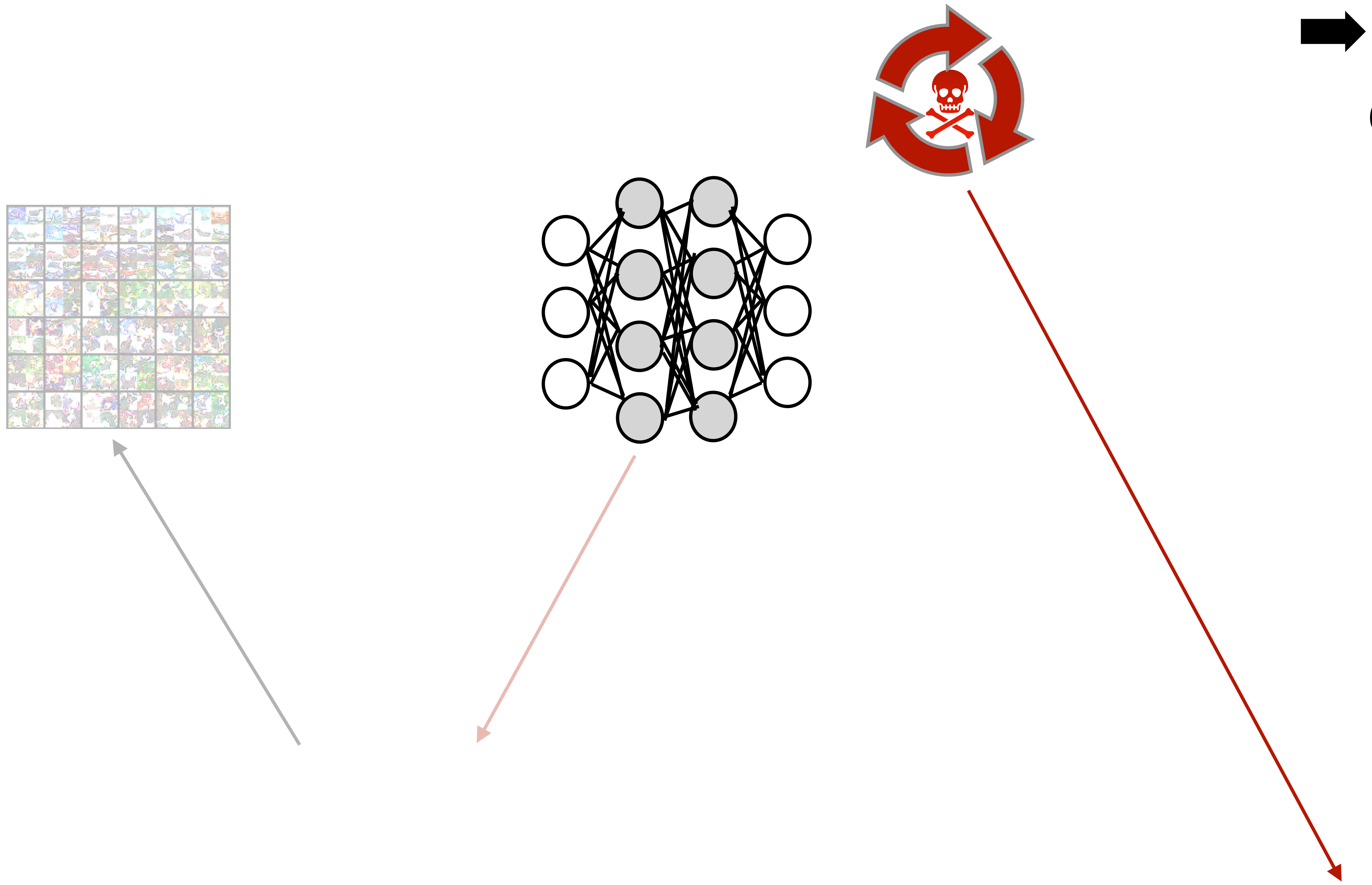
Existing Challenges



Step 1: Train on distilled dataset

Step 2: Retrain on distilled dataset with adversarial perturbations

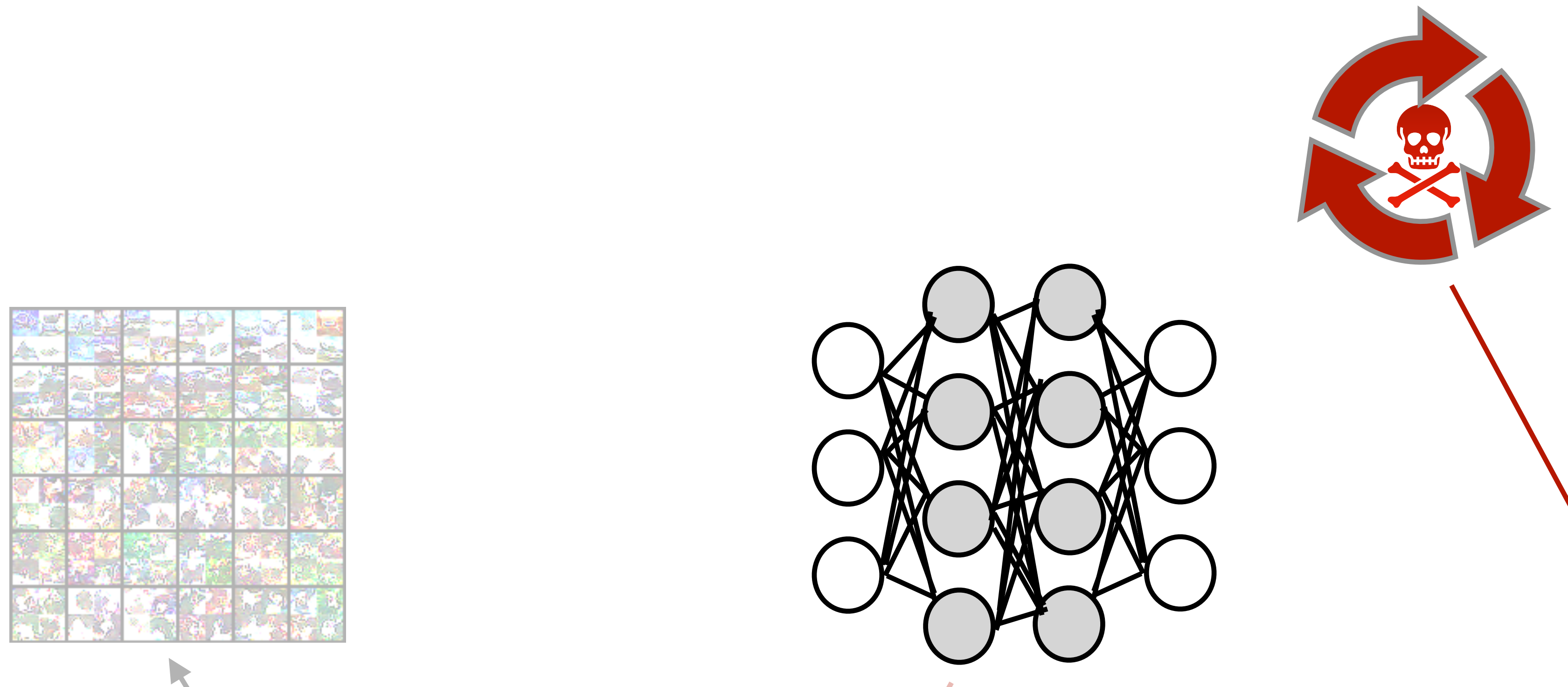
Existing Challenges



➡**High retraining cost**, making the process computationally expensive.

Step 2: Retrain on distilled dataset with adversarial perturbations

Existing Challenges



➡**High retraining cost**, making the process computationally expensive.

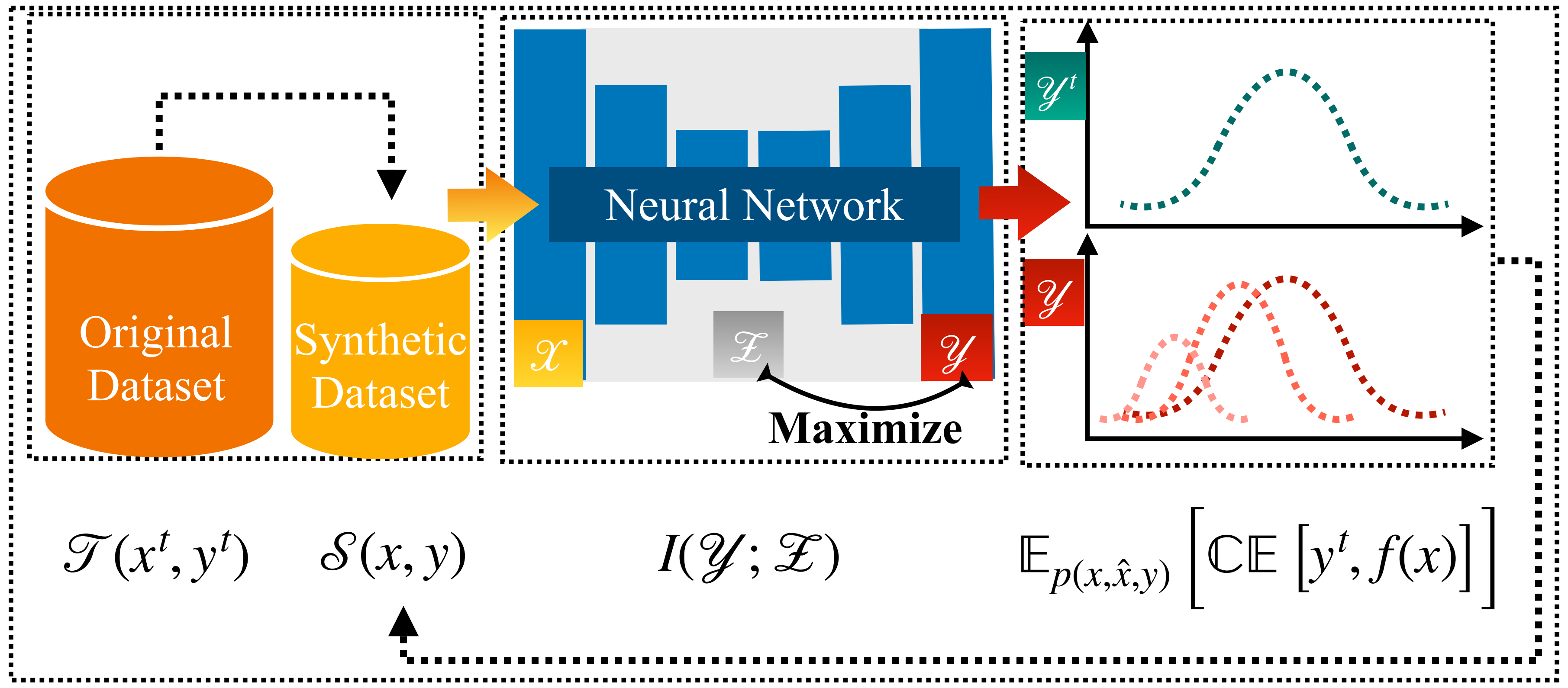
➡**Robustness–accuracy trade-off**, where improving adversarial robustness often reduces clean accuracy.

Step 2: Retrain on distilled dataset with adversarial perturbations

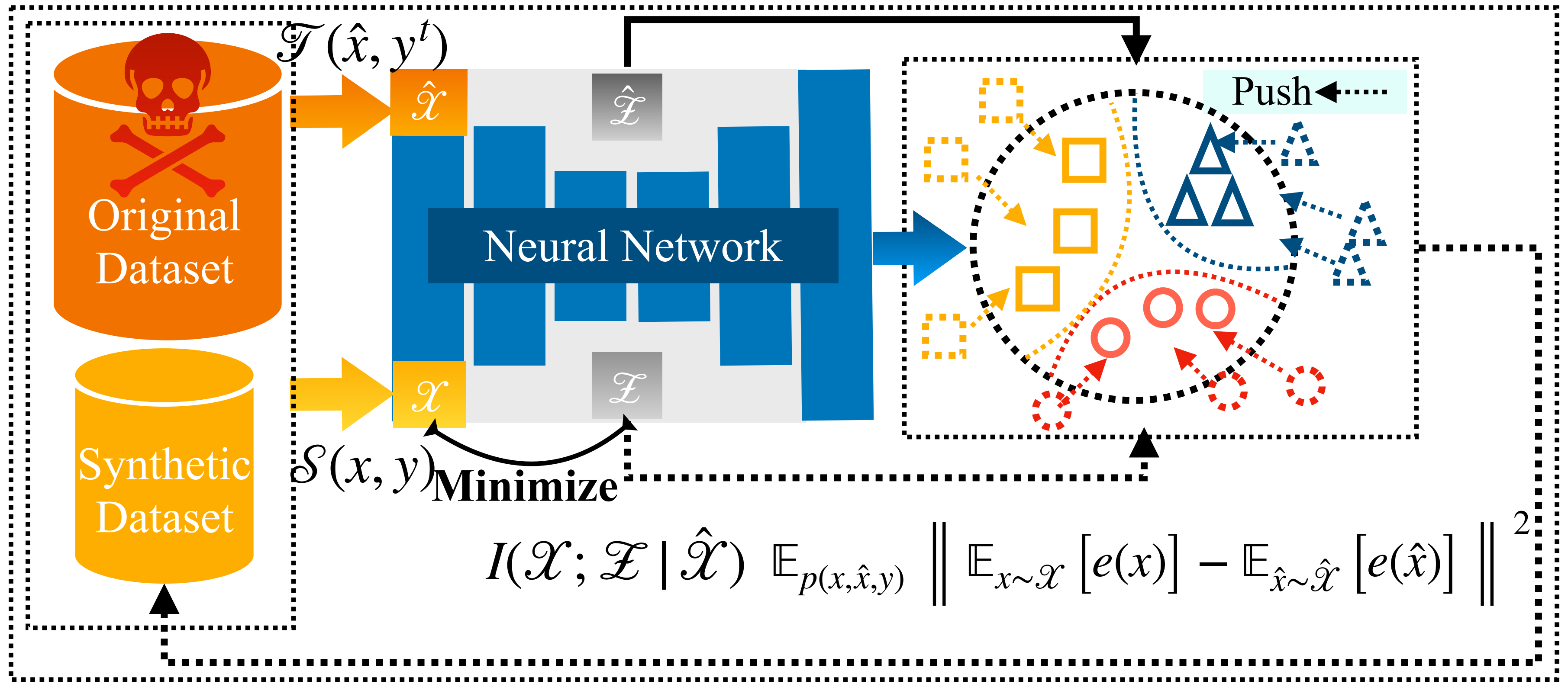
ROME: ROBust distilled datasets via InforMation BottlenEck

ROME: RObust distilled datasets via InforMation BottlenEck

Overview of ROME



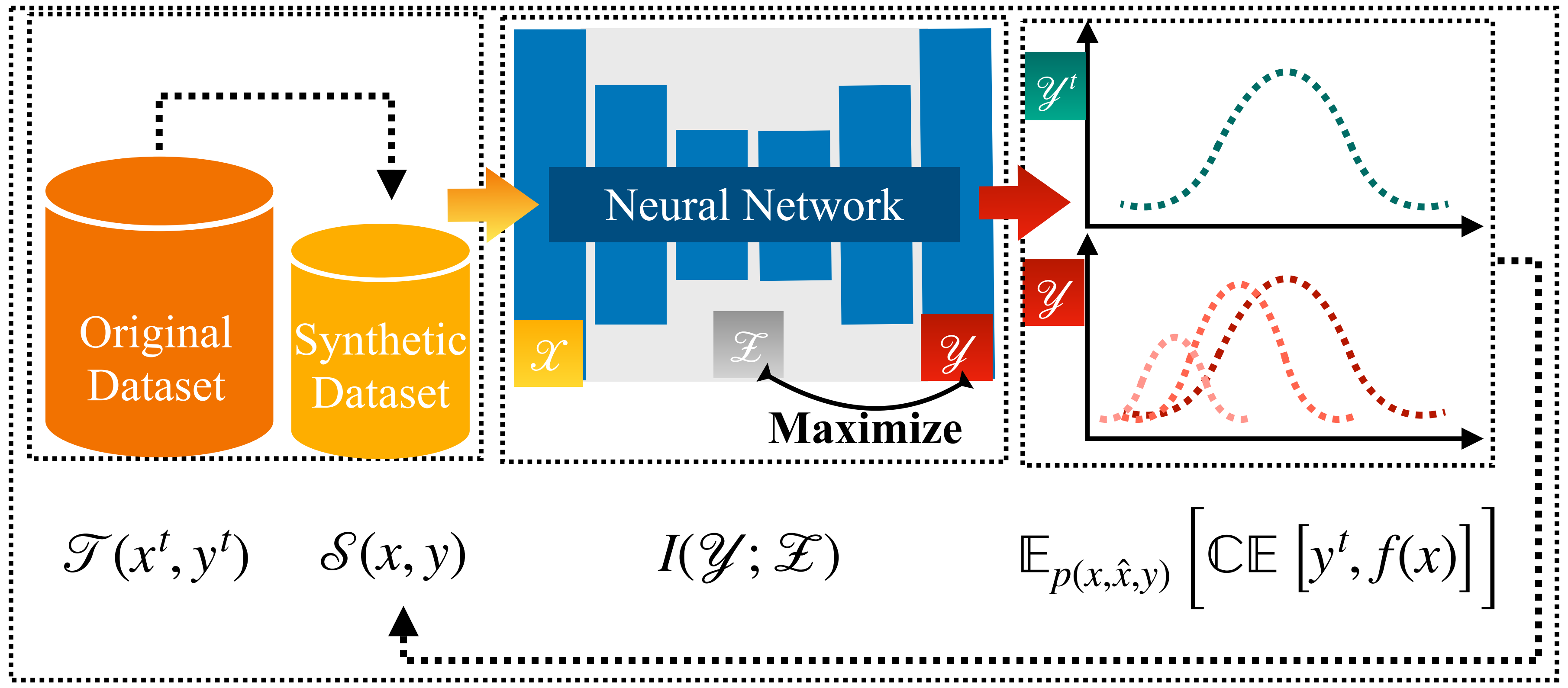
(a) Performance-aligned Term



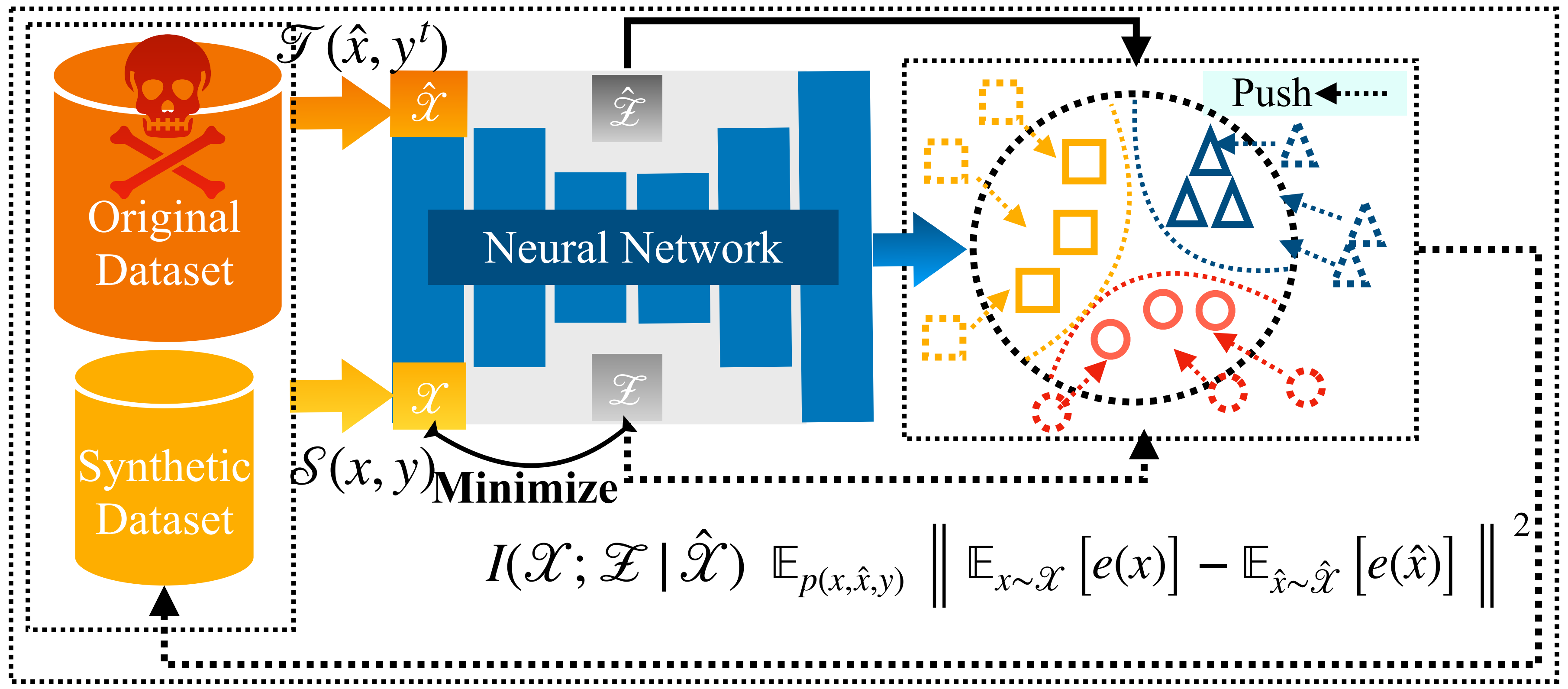
(b) Robustness-aligned Term

ROME: RObust distilled datasets via InforMation BottlenEck

Overview of ROME



(a) Performance-aligned Term



(b) Robustness-aligned Term

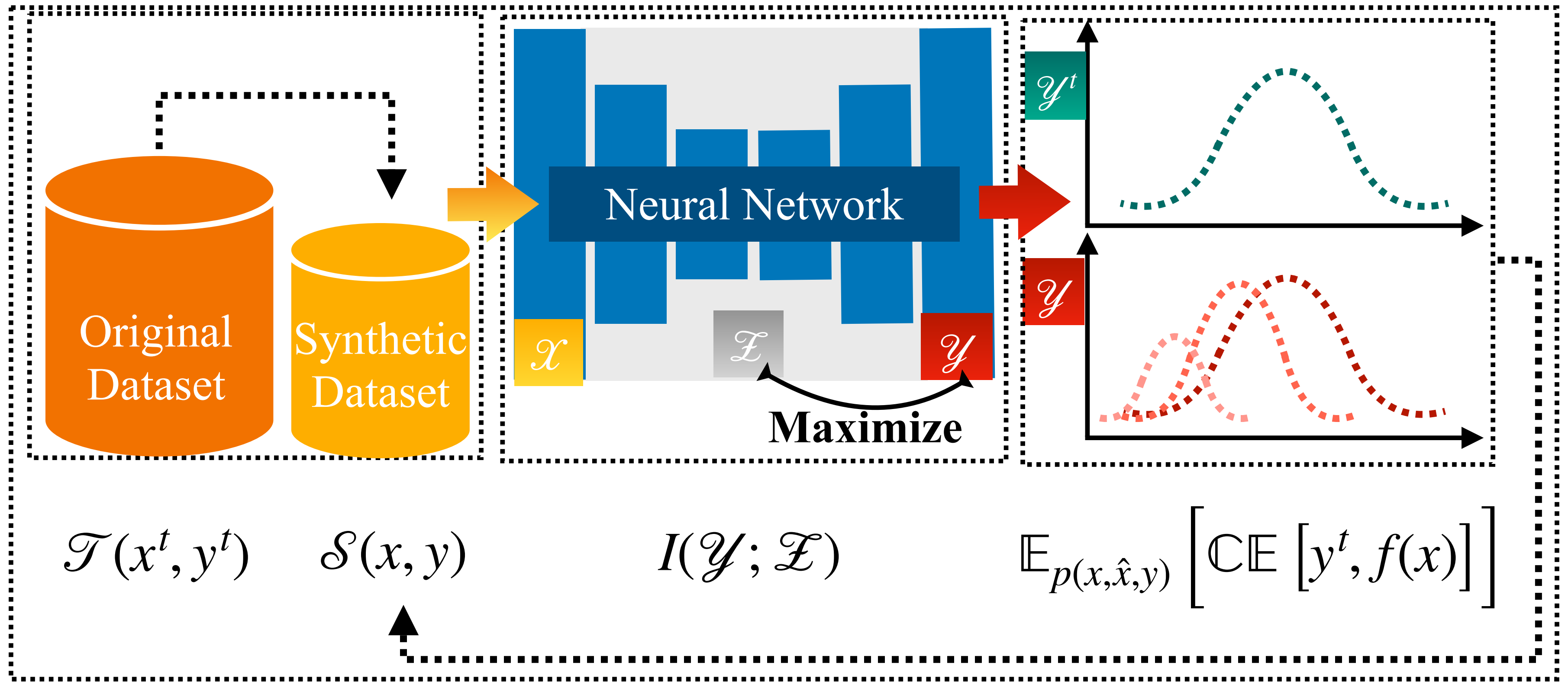
Formulating ROME via information bottleneck

$$\text{ROME} = I(\mathcal{Y}; \mathcal{Z}) - \beta I(\mathcal{X}; \mathcal{Z} \mid \hat{\mathcal{X}})$$

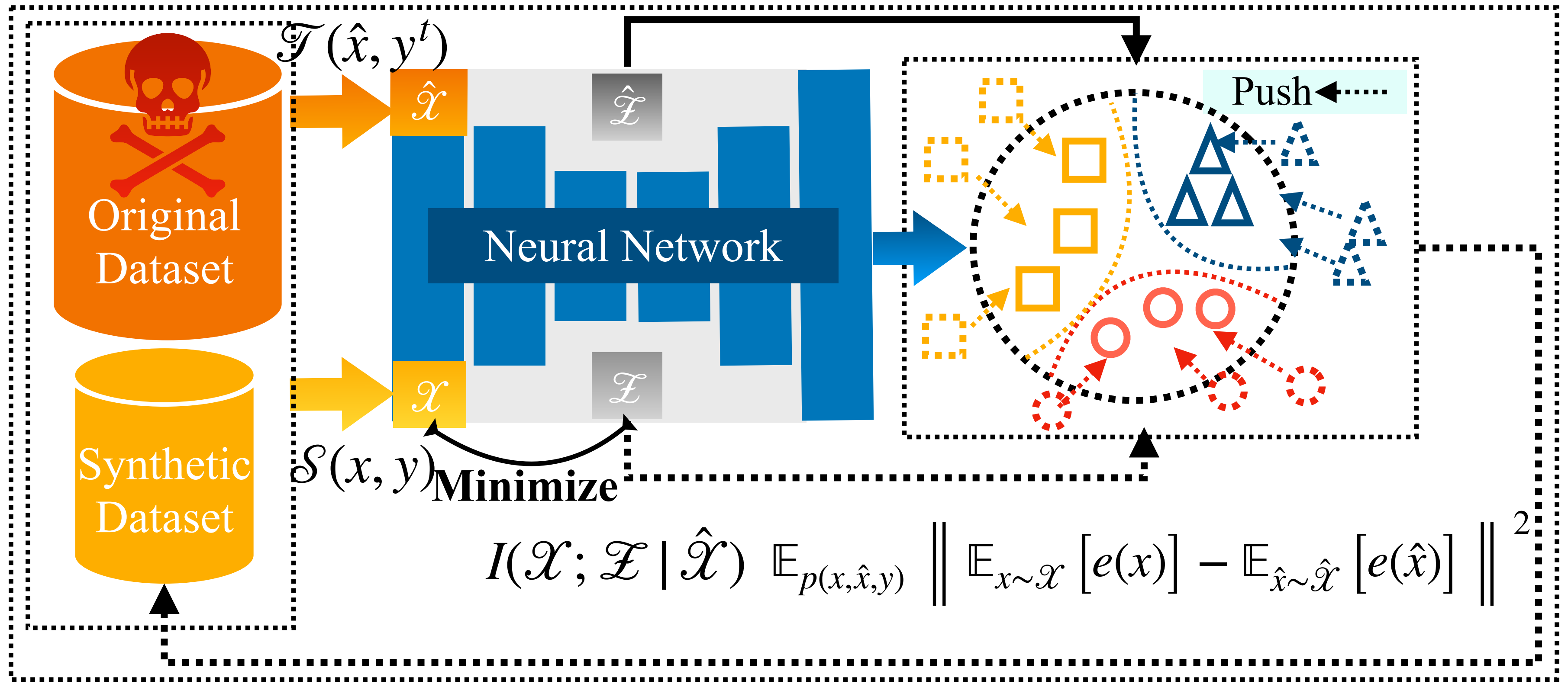
$$\geq \mathbb{E}_{p(x,\hat{x},y)p(z|x,\hat{x},y)}\left[\log q(y \mid z) - \beta \log \frac{p(z \mid x)}{q(z \mid \hat{x})}\right]$$

ROME: RObust distilled datasets via InforMation BottlenEck

Overview of ROME



(a) Performance-aligned Term



(b) Robustness-aligned Term

Formulating ROME via information bottleneck

$$\text{ROME} = I(\mathcal{Y}; \mathcal{Z}) - \beta I(\mathcal{X}; \mathcal{Z} \mid \hat{\mathcal{X}})$$

$$\geq \mathbb{E}_{p(x,\hat{x},y)p(z|x,\hat{x},y)}\left[\log q(y \mid z) - \beta \log \frac{p(z \mid x)}{q(z \mid \hat{x})}\right]$$

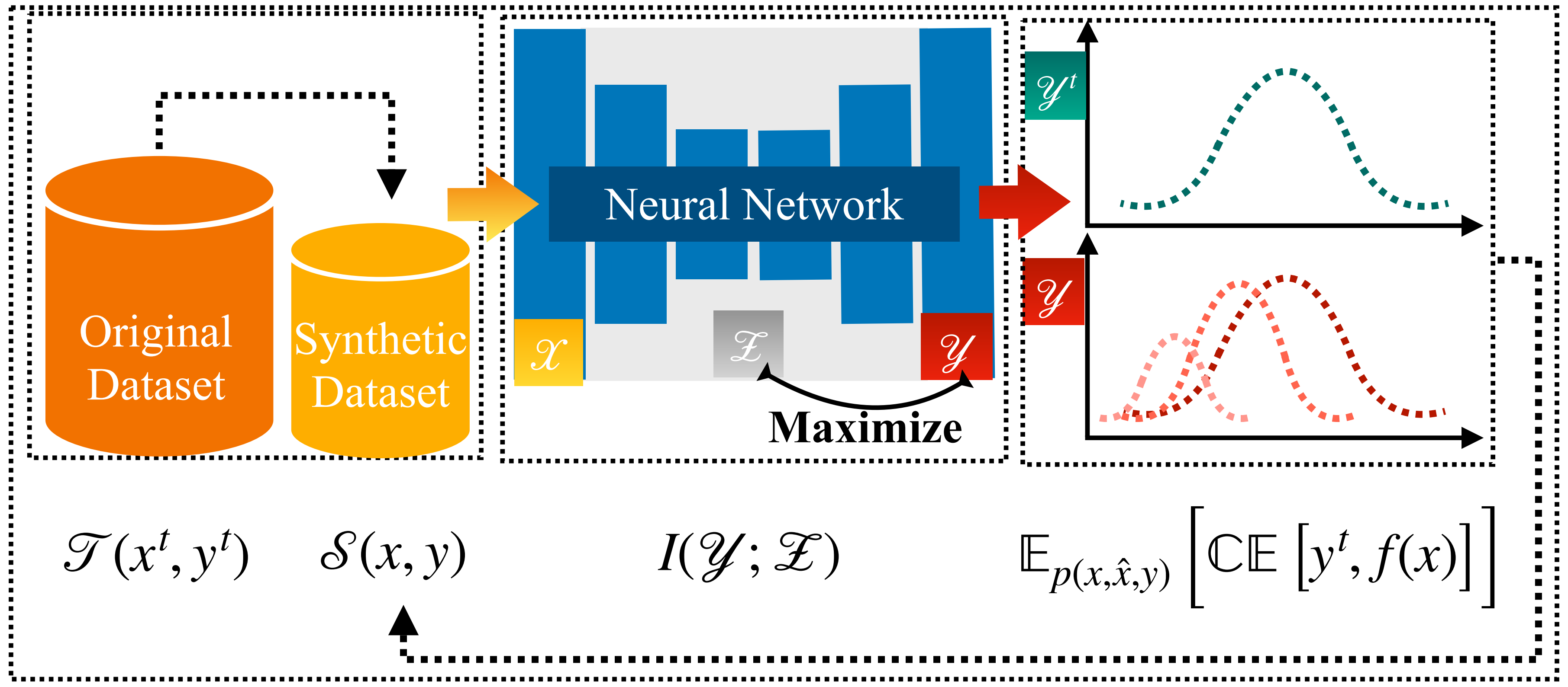
Performance-aligned term

$$\mathcal{L}_{\text{Perf\_Alig}} = \mathbb{E}_{p(x,\hat{x},y)p(z|x,\hat{x},y)}\left[\log q(y \mid z)\right]$$

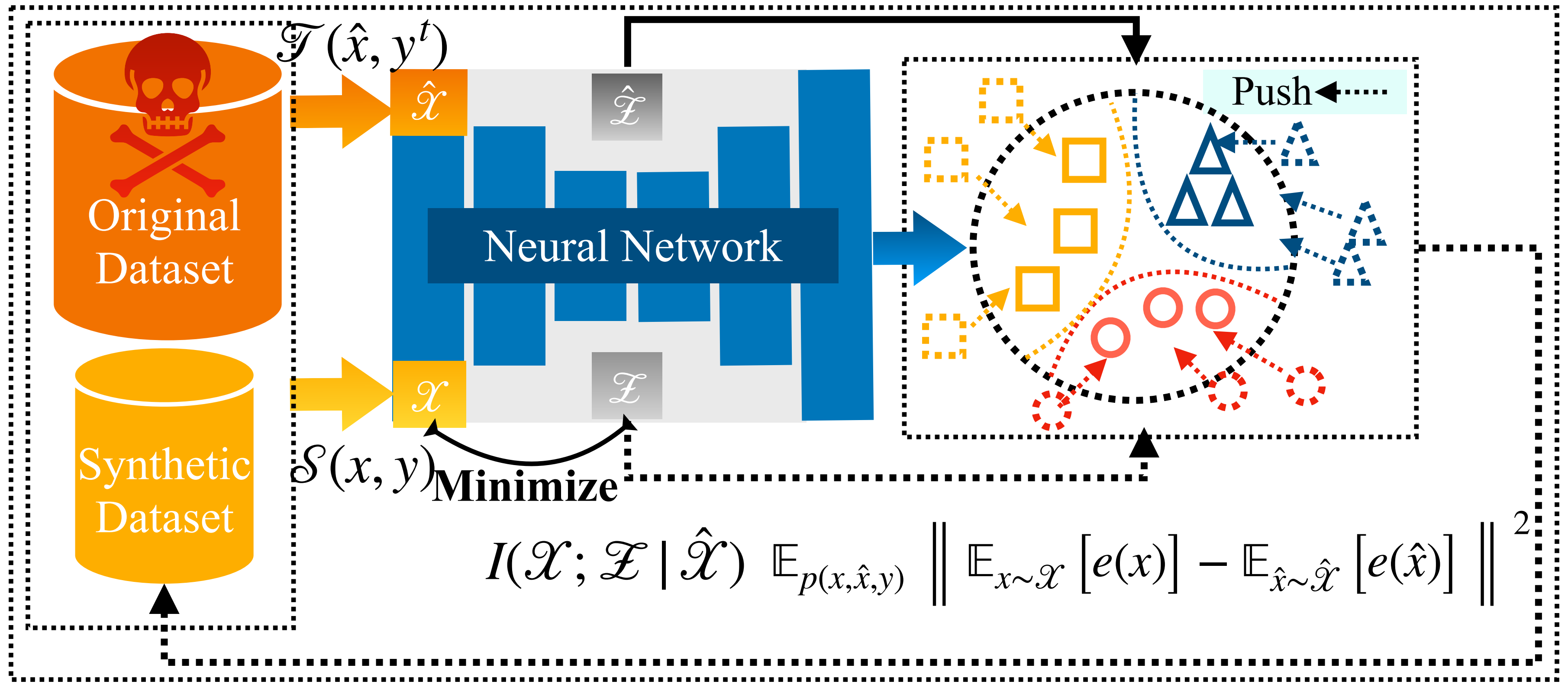
$$= \mathbb{E}_{p(x,\hat{x},y)}\left[\mathbb{CE}\left[y^t, f(x)\right]\right]$$

ROME: ROBust distilled datasets via InforMation BottlenEck

Overview of ROME



(a) Performance-aligned Term



(b) Robustness-aligned Term

Formulating ROME via information bottleneck

$$\text{ROME} = I(\mathcal{Y}; \mathcal{Z}) - \beta I(\mathcal{X}; \mathcal{Z} \mid \hat{\mathcal{X}})$$

$$\geq \mathbb{E}_{p(x,\hat{x},y)p(z|x,\hat{x},y)}\left[\log q(y \mid z) - \beta \log \frac{p(z \mid x)}{q(z \mid \hat{x})}\right]$$

Performance-aligned term

$$\mathcal{L}_{\text{Perf\_Alig}} = \mathbb{E}_{p(x,\hat{x},y)p(z|x,\hat{x},y)}\left[\log q(y \mid z)\right]$$

$$= \mathbb{E}_{p(x,\hat{x},y)}\left[\mathbb{CE}\left[y^t, f(x)\right]\right]$$

Robustness-aligned term

$$\mathcal{L}_{\text{Rob\_Alig}} = \mathbb{E}_{p(x,\hat{x},y)p(z|x,\hat{x},y)}\left[\beta \log \frac{p(z \mid x)}{q(z \mid \hat{x})}\right]$$

$$= \mathbb{E}_{p(x,\hat{x},y)} \left\| \mathbb{E}_{x\sim\mathcal{X}}\left[e(x)\right] - \mathbb{E}_{\hat{x}\sim\hat{\mathcal{X}}}\left[e(\hat{x})\right] \right\|^2$$

Experimental Results

The adversarial robustness of ROME and other dataset distillation methods is evaluated under white-box attack settings.

Table 1. Comparison of model robustness when trained using various DD methods with IPC settings of {1, 10, 50}, against both white-box targeted and untargeted attacks on the CIFAR-10 and CIFAR-100 datasets. Robustness evaluation metrics include RR and CREI, as well as their improved versions I-RR and I-CREI. The best results between the baseline and proposed methods are highlighted in **bold**, while the second-best results are underlined. Improvements in metrics compared to the second-best results are highlighted in red.

Dataset	Method	Targeted Attack				Untargeted Attack			
		RR	CREI	I-RR	I-CREI	RR	CREI	I-RR	I-CREI
CIFAR-10	Full-size	20.42%	24.98%	67.24%	48.39%	28.33%	25.12%	28.82%	25.36%
	DC 2020	30.79%	29.35%	88.51%	58.21%	31.87%	26.70%	56.02%	38.78%
	DSA 2021	45.22%	36.43%	86.81%	57.22%	36.53%	27.75%	53.66%	36.32%
	MTT 2022	36.00%	32.26%	83.95%	56.24%	33.30%	26.26%	48.34%	33.77%
	DM 2023	46.01%	36.01%	85.76%	55.89%	34.50%	28.32%	56.19%	39.16%
	IDM 2023	32.35%	27.75%	87.07%	55.11%	33.03%	28.46%	53.43%	38.66%
	BACON 2024	36.83%	33.05%	84.37%	56.82%	32.87%	27.20%	50.49%	36.01%
	ROME	**81.36%** (35.35 ↑)	**55.28%** (18.85 ↑)	**97.44%** (8.93 ↑)	**63.32%** (5.11 ↑)	**49.86%** (13.33 ↑)	**35.05%** (6.59 ↑)	**67.01%** (10.82 ↑)	**43.62%** (4.46 ↑)
CIFAR-100	Full-size	6.77%	18.18%	65.50%	47.55%	19.91%	18.60%	20.08%	18.69%
	DC 2020	33.11%	30.31%	77.14%	52.32%	28.74%	22.40%	32.33%	24.19%
	DSA 2021	43.97%	35.01%	72.97%	49.51%	28.53%	20.40%	33.29%	22.77%
	MTT 2022	36.06%	31.16%	74.54%	50.40%	26.07%	19.65%	31.10%	22.17%
	DM 2023	39.32%	31.32%	71.29%	47.30%	26.72%	19.78%	29.74%	21.28%
	IDM 2023	34.44%	27.16%	74.57%	47.23%	26.28%	20.36%	30.83%	22.63%
	BACON 2024	31.81%	29.78%	69.96%	48.86%	25.26%	19.30%	27.42%	20.38%
	ROME	**103.09%** (59.12 ↑)	**66.18%** (31.17 ↑)	**100.65%** (23.51 ↑)	**64.96%** (12.64 ↑)	**44.10%** (15.36 ↑)	**28.29%** (5.89 ↑)	**46.24%** (12.95 ↑)	**29.36%** (5.17 ↑)

Experimental Results

The adversarial robustness of ROME and other dataset distillation methods is evaluated under black-box attack settings.

Table 2. Comparison of model robustness measured by I-RR for various dataset distillation methods with IPC-50 under targeted and untargeted transfer-based and query-based black-box attacks on CIFAR-10. Best results are in **bold**, second-best <u>underlined</u>, and improvements over the second-best highlighted in red.

Method	Targeted Attack		Untargeted Attack	
	Transfer	Query	Transfer	Query
DC	85.84%	88.71%	83.97%	43.81%
DSA	<u>94.09%</u>	<u>94.95%</u>	<u>92.31%</u>	54.60%
MTT	91.40%	92.76%	89.02%	48.71%
DM	92.22%	93.86%	90.36%	57.53%
IDM	92.17%	94.37%	89.22%	63.23%
BACON	92.46%	94.67%	89.25%	<u>63.26%</u>
ROME	**99.90%** (5.81 ↑)	**99.79%** (4.84 ↑)	**98.44%** (6.13 ↑)	**78.46%** (15.2 ↑)

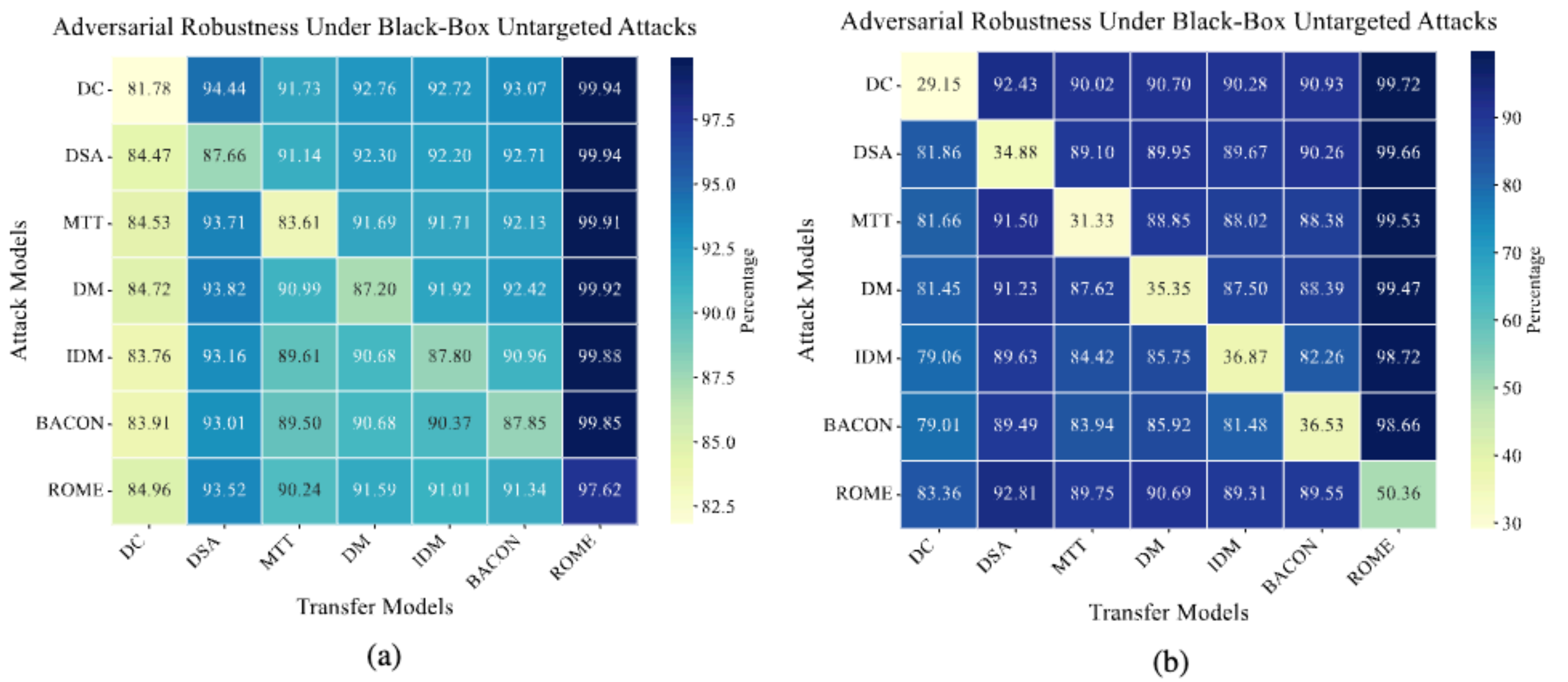


Figure 3. Robustness heatmap of models trained using diverse dataset distillation methods with IPC-50 on CIFAR-10 under targeted and untargeted attacks. The vertical axis represents attacked models, and the horizontal axis shows models used for transfer attacks. Heatmap values represent I-RR, with **darker colors** indicating **higher I-RR** and thus **better robustness** against adversarial attacks.

Experimental Results

The adversarial robustness and training efficiency of ROME and other dataset distillation methods are evaluated.

Table 3. Comparison of adversarial robustness (I-CREI, %) and training time (hours) of ROME and baseline dataset distillation methods on CIFAR-10 (IPC-50) under targeted attacks. "Base" indicates standard distillation training, while "+AdvTrain" refers to the additional time required for adversarial training to improve robustness. Best results, balancing robustness and efficiency, are highlighted in **bold**, and † denotes consistent results from "Base" to "+AdvTrain", indicating no need for adversarial fine-tuning.

Method	I-CREI		Training Time	
	Base	+AdvTrain	Base	+AdvTrain
DC	58.21%	63.43%	0.425	1.088
DSA	57.22%	63.46%	0.437	1.103
MTT	56.24%	62.44%	0.444	1.088
DM	55.89%	63.21%	0.452	1.109
IDM	55.11%	63.11%	0.414	1.055
BACON	56.82%	62.68%	0.442	1.101
ROME	**63.32%**	**63.32%** †	**0.418**	**0.418** †

Experimental Results

Ablation studies are conducted on various configurations, with visualizations illustrating the impact of different hyperparameters.

Table 4. Ablation studies on the Robust Pretrained Model (RPM) and Adversarial Perturbation (AP) under both targeted and untargeted attacks, evaluated by I-RR and I-CREI on the CIFAR-10 dataset with IPC-50. Best results are highlighted in **bold**.

Configuration	Targeted Attack		Untargeted Attack	
	I-RR	I-CREI	I-RR	I-CREI
Baseline	81.86%	55.26%	32.45%	29.29%
+RPM	84.50%	56.53%	34.89%	30.45%
+AP	94.66%	61.67%	47.64%	36.78%
+RPM&AP	**97.73%**	**63.23%**	**51.73%**	**38.95%**

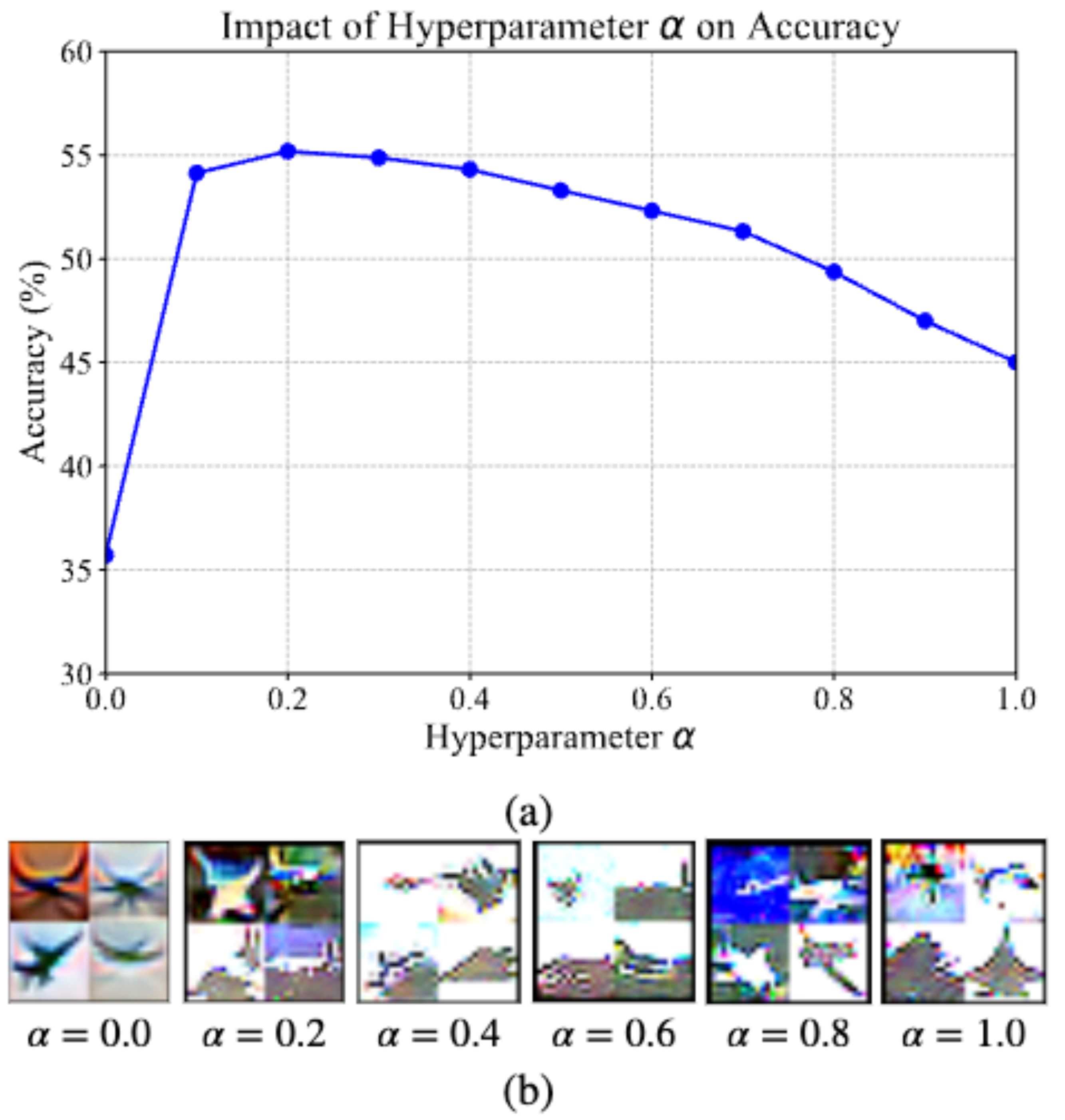


Figure 4. Ablation study of the hyperparameter α. (a) Displays the accuracy (y-axis) as a function of α (x-axis) for different values of α, and (b) shows the corresponding visualizations for these values.

Thank you!

If you're interested in **adversarial robustness** or **dataset distillation**, *feel free to reach out.*

E-mail: zhengzhou@buaa.edu.cn

Personal Website: https://zhouzhengqd.github.io/

Scan the QR codes for more information.

Code

Project Page

Contact us