Centre for Frontier AI Research
CFAR

NEURAL INFORMATION PROCESSING SYSTEMS

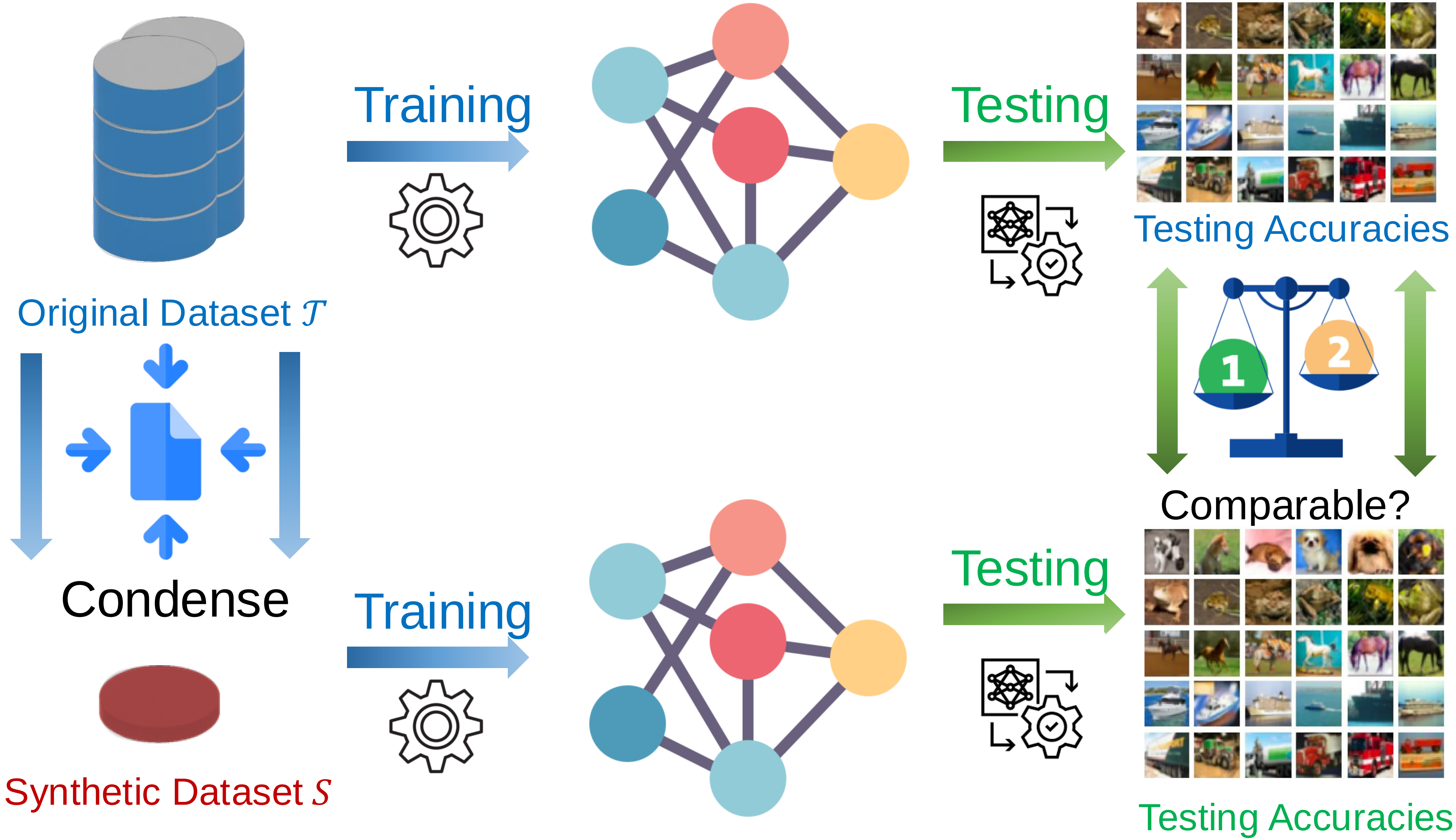
Diversity-Driven Synthesis: Enhancing Dataset Distillation through Directed Weight Adjustment

Jiawei Du, Xin Zhang, Juncheng Hu, Wenxing Huang, Joey Tianyi Zhou

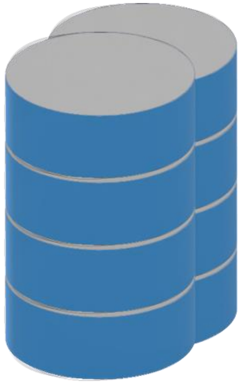
Centre for Frontier AI Research (CFAR, A*STAR)

CREATING GROWTH, ENHANCING LIVES

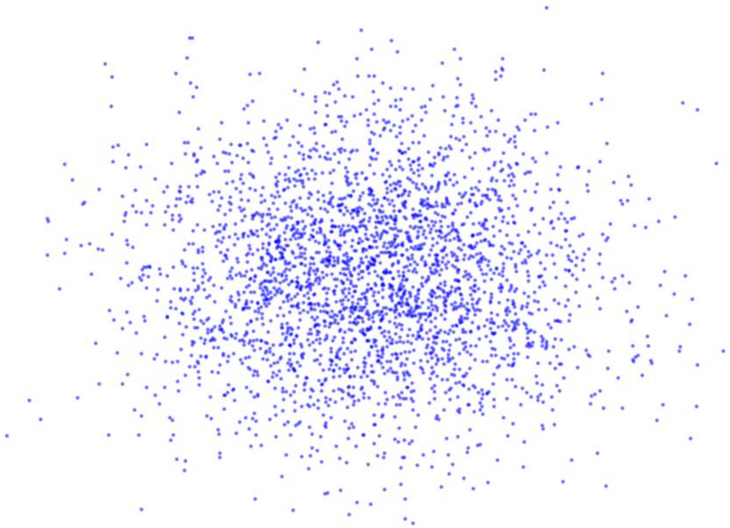
Dataset Distillation (DD)



An intuition

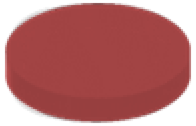


Original Dataset $\mathcal{T}$

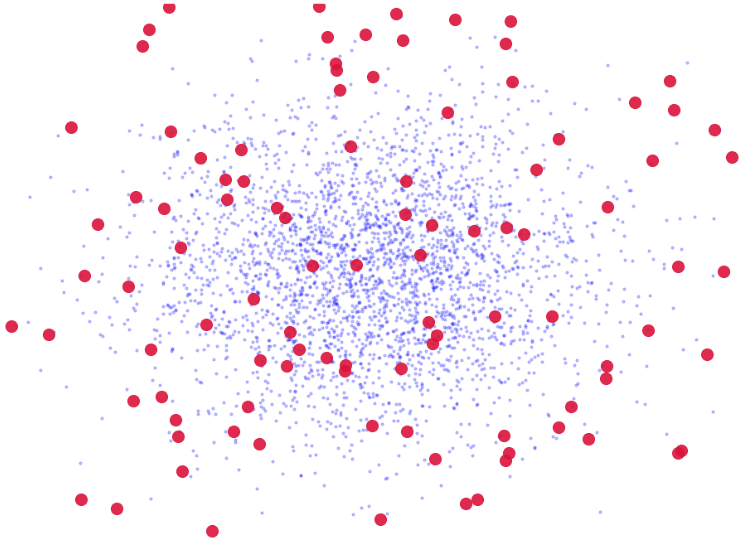


Visualize data samples (3000 points)

- Duplicated easy samples
- Rare hard samples



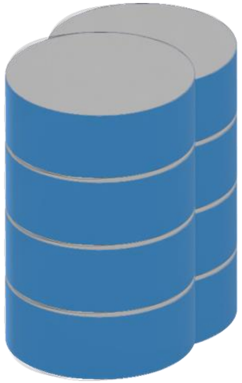
Synthetic Dataset S



Distilled data samples (100 points)

DD aims to use **fewer, deduplicated** samples to represent the entire data space.

Baseline



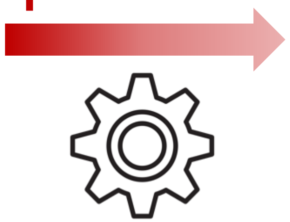
Original Dataset $\mathcal{T}$

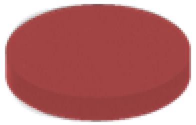
Optimize θ

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left[\ell\left(h_{\theta_{\mathcal{T}}},\boldsymbol{x}\right)\right]$$

Minimized Loss

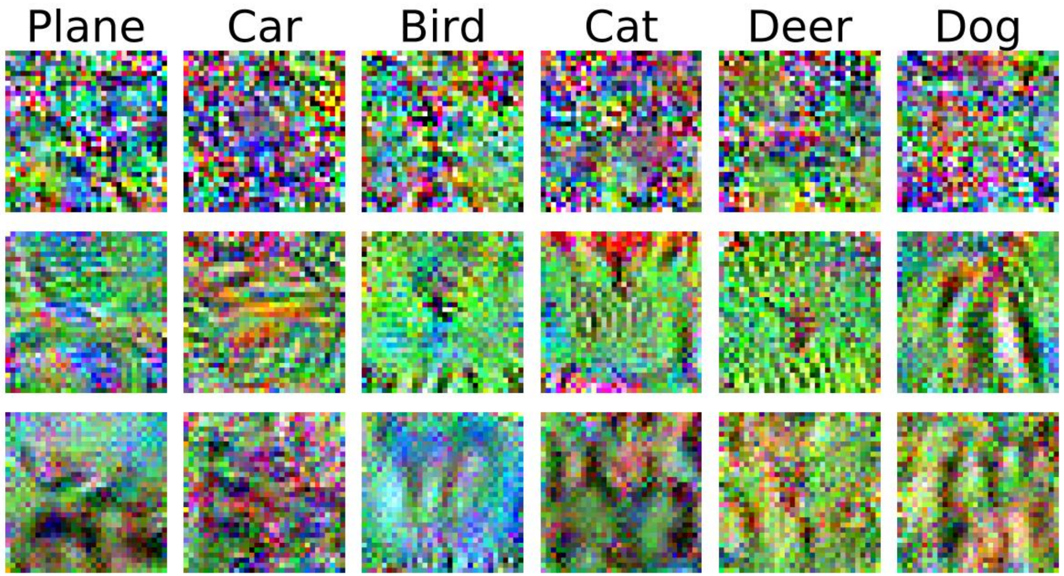
Optimize x





Synthetic Dataset S

$$\mathop{\arg\min}_{\boldsymbol{s}_i\in\mathbb{R}^d}\ell\left(f_{\theta_{\mathcal{T}}},\boldsymbol{s}_i\right)$$



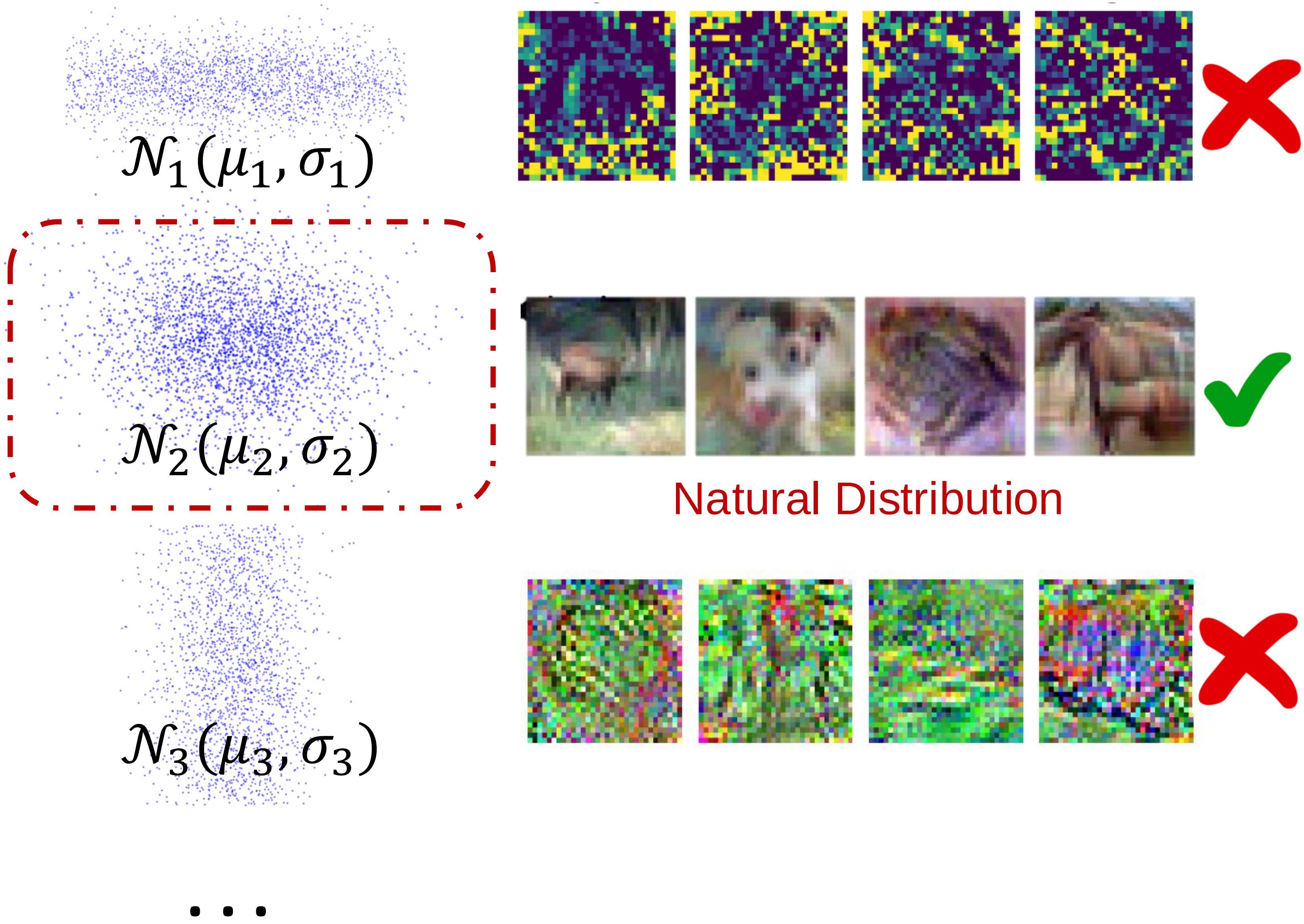
Dataset Distillation (wang et al. 2018)

Poor distillation performance

Solutions with different BN

$$\underset{\boldsymbol{s}_i \in \mathbb{R}^d}{\arg\min}\, \ell\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right)$$

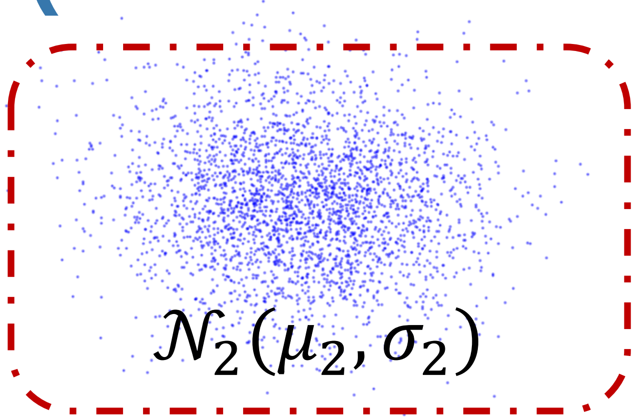
Solve it can obtain



Solutions with different Batch Norm.

Baseline: Sre2L (Yin et al.2023)

To obtain



Natural Distribution

Sre2L proposes to use a BN loss to constrain

$$\mathcal{L}_{\mathrm{BN}} = \mathcal{L}_{\mathrm{mean}} + \mathcal{L}_{\mathrm{var}} \quad \text{where} \quad \mathcal{L}_{\mathrm{mean}}\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right) = \sum_l \left\|\mu_l\left(\mathbb{S}\right) - \mu_l\left(\mathcal{T}\right)\right\|_2,$$

$$\text{and} \quad \mathcal{L}_{\mathrm{var}}\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right) = \sum_l \left\|\sigma_l^2\left(\mathbb{S}\right) - \sigma_l^2\left(\mathcal{T}\right)\right\|_2,$$

Distillation progress is to solve

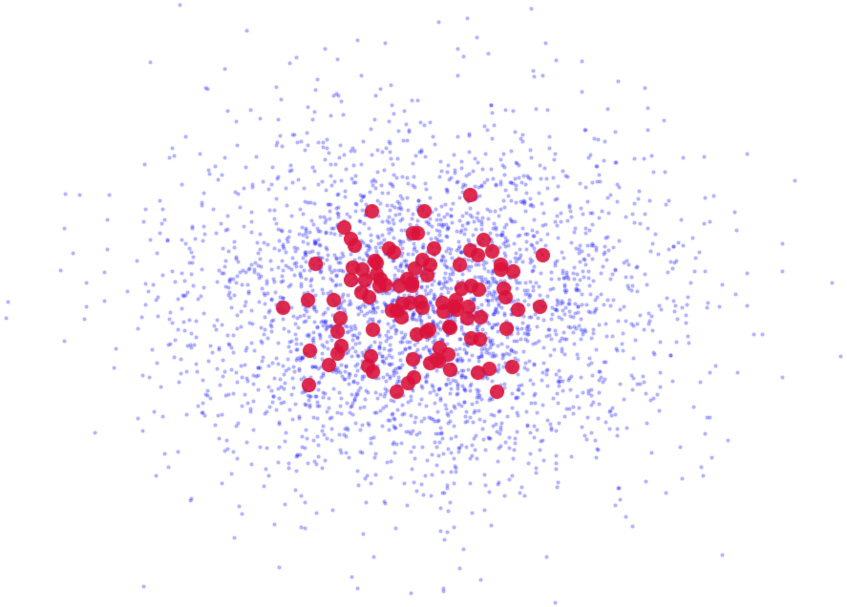
$$\underset{\boldsymbol{s}_i \in \mathbb{R}^d}{\arg\min} \left[\ell\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right) + \lambda \mathcal{L}_{\mathrm{BN}}\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right)\right]$$

Diversity limitations

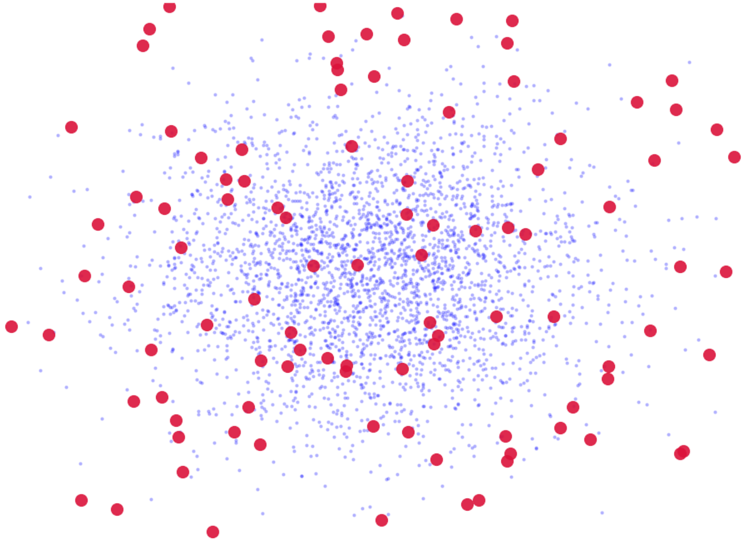
Distillation progress is to solve

$$\underset{\boldsymbol{s}_i \in \mathbb{R}^d}{\arg\min} \left[\ell\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right) + \lambda \mathcal{L}_{\mathrm{BN}}\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right)\right]$$

However, distilled data is clustered at the central



Sre2L distilled data



Ideal distilled data

The Diversity issue motivates our study

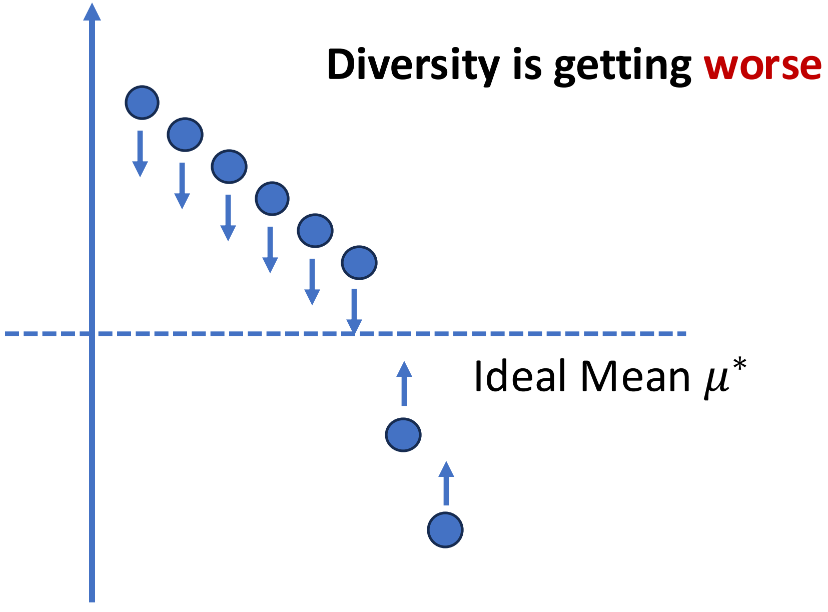
Diversity limitations

The clustering is caused by the contradictory in BN loss

$$\mathcal{L}_{\text{BN}} = \mathcal{L}_{\text{mean}} + \mathcal{L}_{\text{var}} \quad \text{where} \quad \mathcal{L}_{\text{mean}}\left(f_{\theta_\mathcal{T}}, \boldsymbol{s}_i\right) = \sum\nolimits_l \left\|\mu_l\left(\mathbb{S}\right) - \mu_l\left(\mathcal{T}\right)\right\|_2,$$

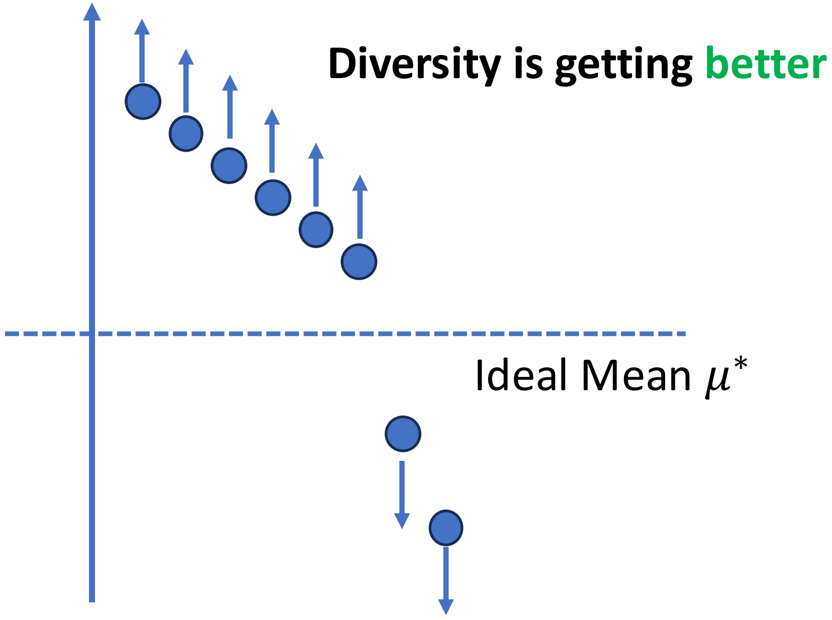
$$\text{and} \quad \mathcal{L}_{\text{var}}\left(f_{\theta_\mathcal{T}}, \boldsymbol{s}_i\right) = \sum\nolimits_l \left\|\sigma_l^2\left(\mathbb{S}\right) - \sigma_l^2\left(\mathcal{T}\right)\right\|_2,$$

For simplicity, we discuss the minimization in the 1D case



If we minimize $\mathcal{L}_{mean}$ only

In contrast



If we minimize $\mathcal{L}_{var}$ only

Diversity limitations

The clustering is caused by the contradictory in BN loss, theoretical proof is provided

For $\frac{\partial \mathcal{L}_{\text{mean}}}{\partial \boldsymbol{s}_i}$, we have

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{mean}}}{\partial \boldsymbol{s}_i} = \frac{\partial \left[\mu(\mathcal{S}) - \mu(\mathcal{T})\right]^2}{\partial \boldsymbol{s}_i} &= \frac{\partial \left[\mu(\mathcal{S}) - \mu(\mathcal{T})\right]^2}{\partial \mu(\mathcal{S})} \cdot \frac{\partial \mu(\mathcal{S})}{\partial \boldsymbol{s}_i} \\ &= 2\left[\mu(\mathcal{S}) - \mu(\mathcal{T})\right] \cdot \frac{1}{|\mathcal{S}|},\end{aligned}$$

because $\mu(\mathcal{S}) = \frac{1}{|\mathcal{S}|}\boldsymbol{s}_i + \sum_{j \neq i} \frac{1}{|\mathcal{S}|}\boldsymbol{s}_j$, thus $\frac{\partial \mu(\mathcal{S})}{\partial \boldsymbol{s}_i} = \frac{1}{|\mathcal{S}|}$. For $\frac{\partial \mathcal{L}_{\text{var}}}{\partial \boldsymbol{s}_i}$, we have

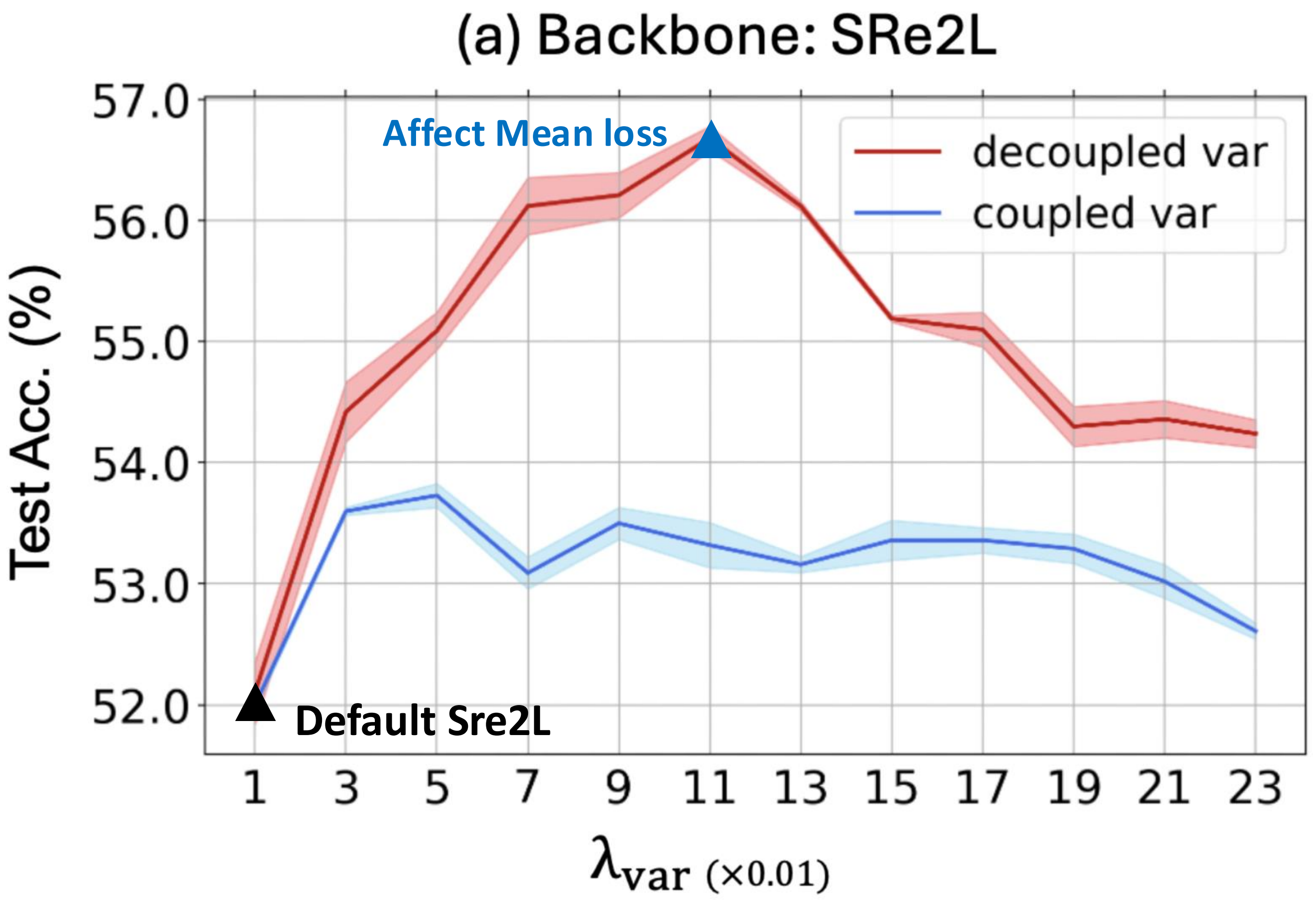
$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{var}}}{\partial \boldsymbol{s}_i} = \frac{\partial \left[\sigma^2(\mathcal{S}) - \sigma^2(\mathcal{T})\right]^2}{\partial \boldsymbol{s}_i} &= \frac{\partial \left[\sigma^2(\mathcal{S}) - \sigma^2(\mathcal{T})\right]^2}{\partial \sigma^2(\mathcal{S})} \cdot \frac{\partial \sigma^2(\mathcal{S})}{\partial \boldsymbol{s}_i} \\ &= 2\left[\sigma^2(\mathcal{S}) - \sigma^2(\mathcal{T})\right] \cdot \frac{\partial \sigma^2(\mathcal{S})}{\partial \boldsymbol{s}_i} \\ &= 2\left[\sigma^2(\mathcal{S}) - \sigma^2(\mathcal{T})\right] \cdot \frac{\partial \left[\frac{1}{|\mathcal{S}|}\left(\boldsymbol{s}_i - \mu(\mathcal{S})\right)^2 + \sum_{j \neq i} \frac{1}{|\mathcal{S}|}\left(\boldsymbol{s}_j - \mu(\mathcal{S})\right)^2\right]}{\partial \boldsymbol{s}_i} \\ &= 2\left[\sigma^2(\mathcal{S}) - \sigma^2(\mathcal{T})\right] \cdot \frac{1}{|\mathcal{S}|} \frac{\partial \left(\boldsymbol{s}_i - \mu(\mathcal{S})\right)^2}{\partial \boldsymbol{s}_i} \\ &= 2\left[\sigma^2(\mathcal{S}) - \sigma^2(\mathcal{T})\right] \cdot \frac{1}{|\mathcal{S}|} \cdot 2\left(\boldsymbol{s}_i - \mu(\mathcal{S})\right) \cdot \frac{\partial \left(\boldsymbol{s}_i - \mu(\mathcal{S})\right)}{\partial \boldsymbol{s}_i} \\ &= 2\left[\sigma^2(\mathcal{S}) - \sigma^2(\mathcal{T})\right] \cdot \frac{1}{|\mathcal{S}|} \cdot 2\left(\boldsymbol{s}_i - \mu(\mathcal{S})\right) \cdot \left(1 - \frac{1}{|\mathcal{S}|}\right).\end{aligned}$$

CREATING GROWTH, ENHANCING LIVES

Our feasibility experiments

We decouple the BN loss and **emphasize** the variation loss only

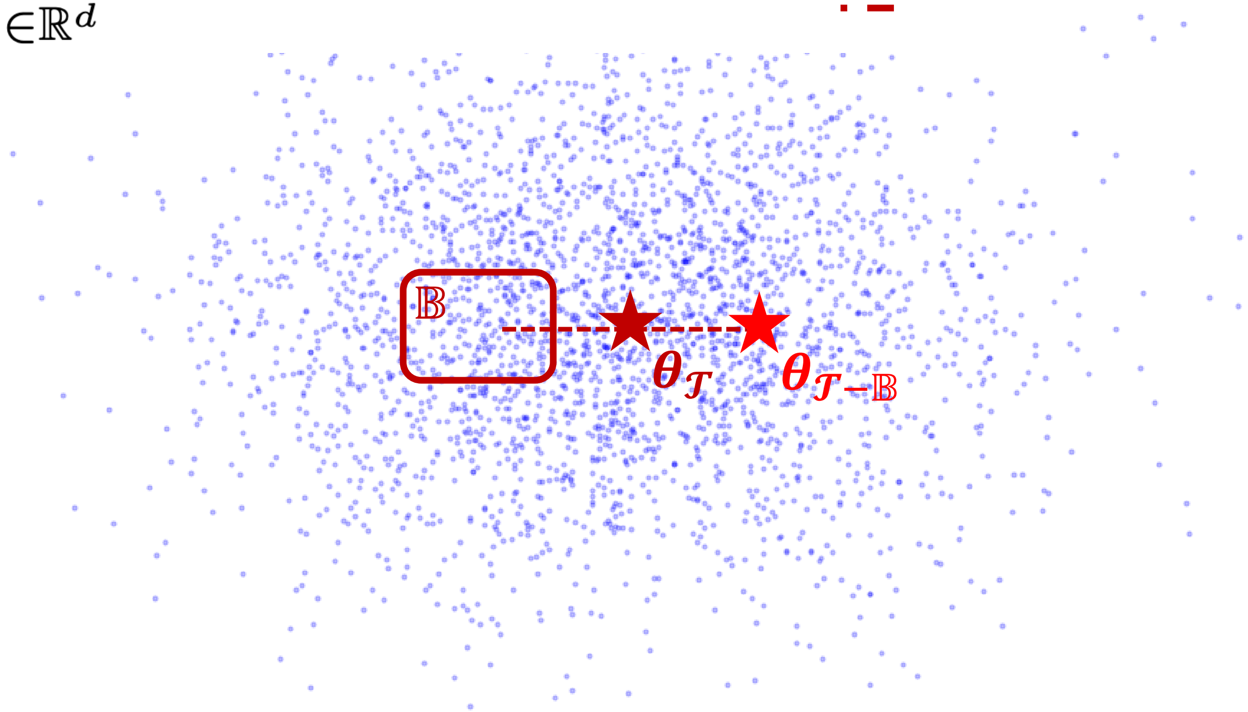
$$\mathcal{L}_{\text{mean}}\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right) + \boxed{\lambda_{\text{var}}} \mathcal{L}_{\text{var}}\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right)$$



Directed Weight Adjustment (DWA)

We move further to perturb distillation progress for enhanced diversity

$$\underset{\boldsymbol{s}_i \in \mathbb{R}^d}{\arg\min} \left[\ell\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right) + \lambda \mathcal{L}_{\mathrm{BN}}\left(f_{\theta_{\mathcal{T}} \boxed{+\epsilon}}, \boldsymbol{s}_i\right)\right]$$

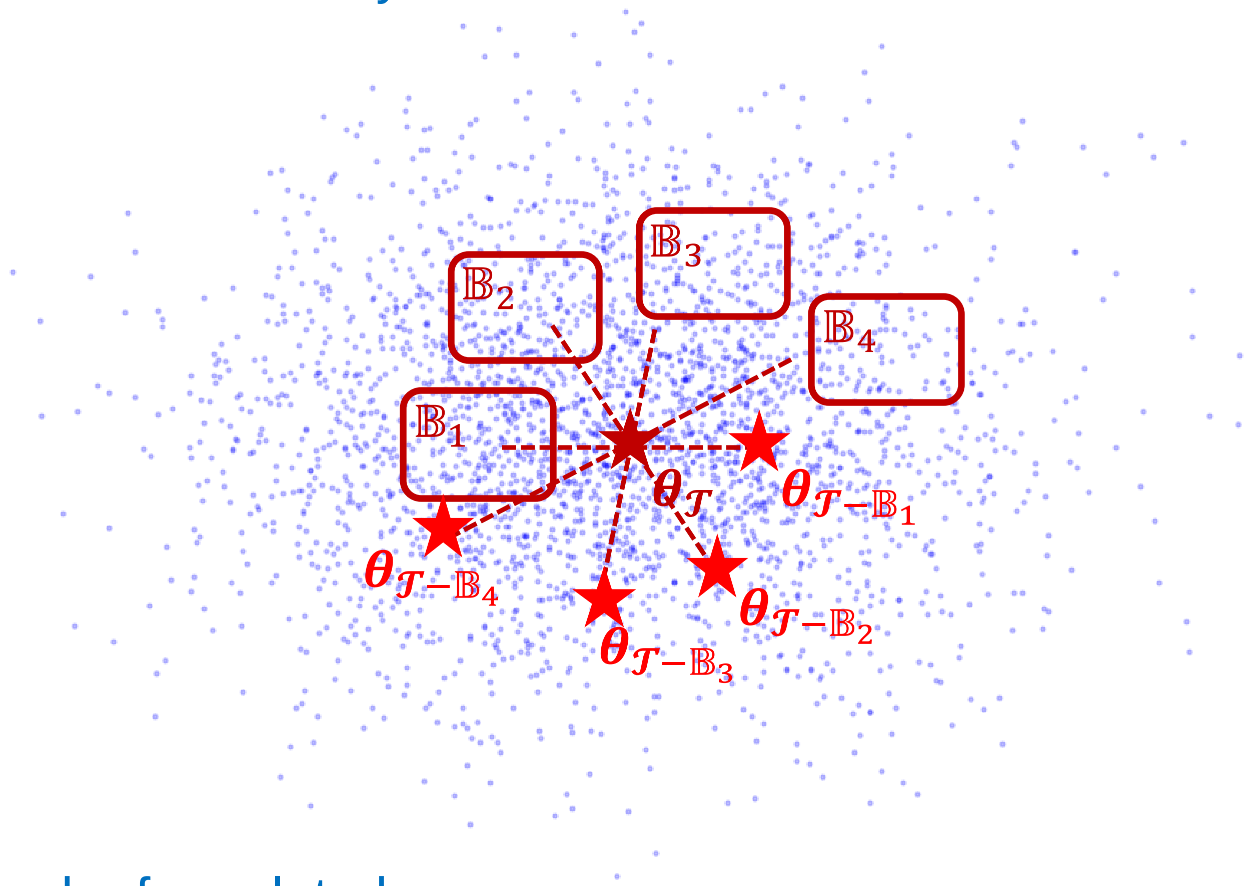


- We perturb $\theta_{\mathcal{T}}$ carefully to make it represent the majority of $\mathcal{T}$
- The majority is $\mathcal{T}/\mathbb{B}$, which removes a random set $\mathbb{B}$ from $\mathcal{T}$

$$\theta_{\mathcal{T}\setminus\mathbb{B}} = \theta_{\mathcal{T}} + \nabla_{\theta} L_{\mathbb{B}}\left(f_{\theta_{\mathcal{T}}}\right)$$

Directed Weight Adjustment (DWA)

If we sample the random set many times



Our DWA can thus be formulated as

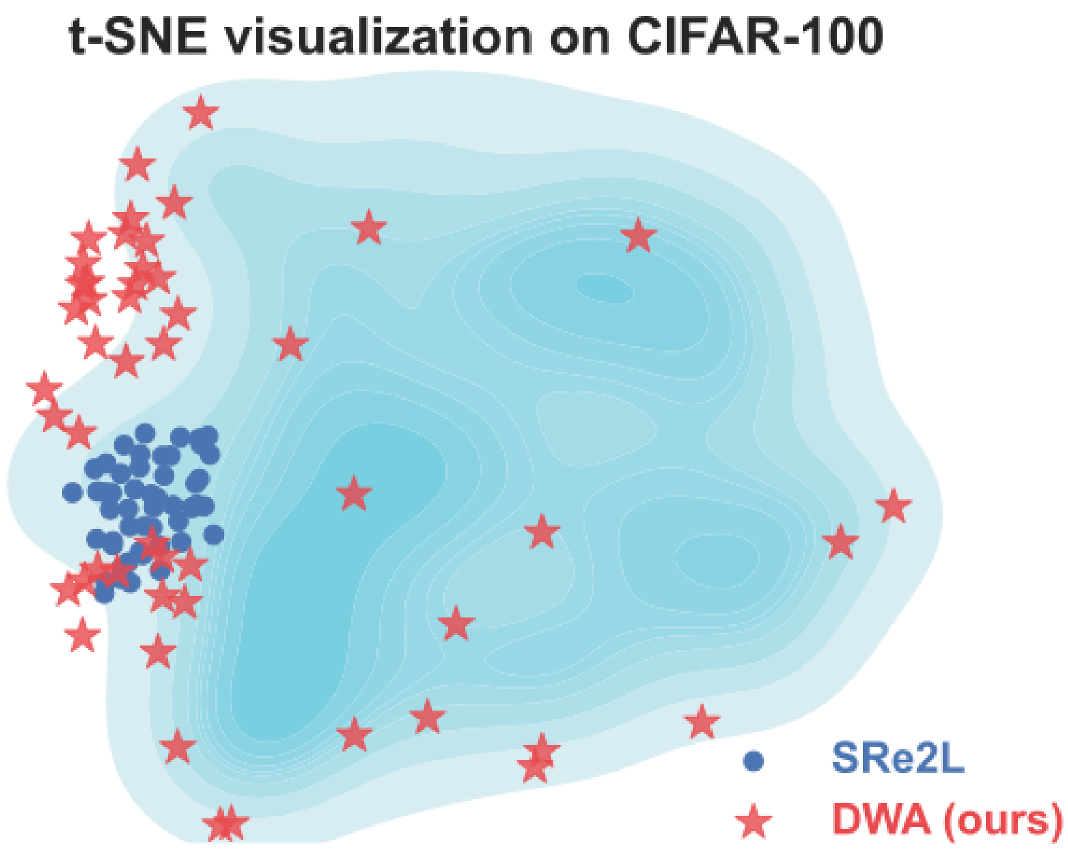
$$\tilde{\boldsymbol{s}}_i = \underset{\boldsymbol{s}\in\mathbb{R}^d}{\arg\min}\,\mathcal{L} \quad \text{where} \quad \mathcal{L} = \left[\ell\left(f_{\theta_{\mathcal{T}}+\widetilde{\Delta\theta}}, \boldsymbol{s}_i\right) + \lambda\mathcal{L}_{\text{mean}}\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right) + \lambda_{\text{var}}\mathcal{L}_{\text{var}}\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right)\right]$$

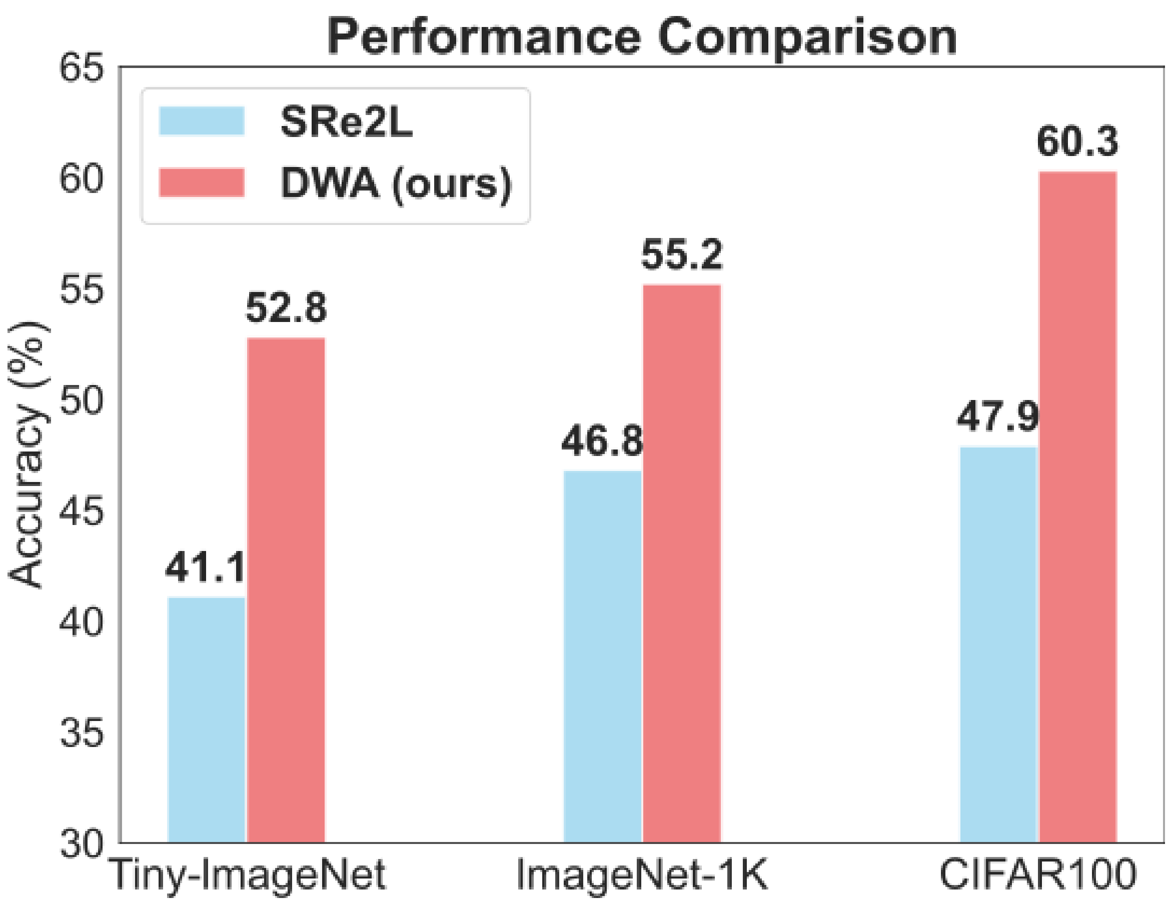
$$\widetilde{\Delta\theta} = \underset{\Delta\theta}{\arg\max}\, L_{\mathbb{B}}\left(f_{\theta_{\mathcal{T}}+\Delta\theta}\right) \quad \text{where} \quad L_{\mathbb{B}}\left(f_{\theta_{\mathcal{T}}+\Delta\theta}\right) = \sum_{\boldsymbol{x}_i\in\mathbb{B}} \ell\left(f_{\theta_{\mathcal{T}}+\Delta\theta}, \boldsymbol{x}_i\right),$$

We could enhance diversity but do not introduce noise

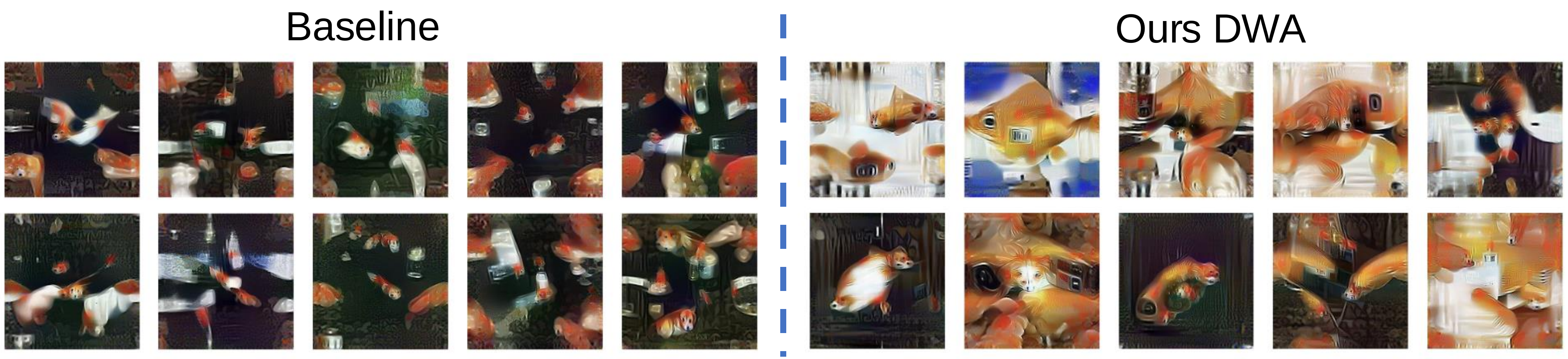
Experiments

We could enhance diversity but do not introduce noise





Experiments



Visualization of distilled images of ImageNet-1k

Dataset	ipc	ResNet-18		ResNet-50		ResNet-101	
		SRe2L [46]	DWA (ours)	SRe2L	DWA (ours)	SRe2L	DWA (ours)
Tiny-ImageNet	50	41.1±0.4	52.8±0.2	42.2±0.5	53.7±0.2	42.5±0.2	54.7±0.3
	100	49.7±0.3	56.0±0.2	51.2±0.4	56.9±0.4	51.5±0.3	57.4±0.3
ImageNet-1K	10	21.3±0.6	37.9±0.2	28.4±0.1	43.0±0.5	30.9±0.1	46.9±0.4
	50	46.8±0.2	55.2±0.2	55.6±0.3	62.3±0.1	60.8±0.5	63.3±0.7
	100	52.8±0.3	59.2±0.3	61.0±0.4	65.7±0.4	62.8±0.2	66.7±0.2

Results in ImageNet-1k

Theoretical proof and more experimental results can be found in our paper

arXiv

Github