Diversity-Enhanced Distribution Alignment for Dataset Distillation

Supplementary Material

8. Experimental Setup

We follow the experimental setup of SRe2L [43]. The detailed parameter settings are documented in the following tables: CIFAR-10/100 in Table 5, Tiny-ImageNet in Table 6, and ImageNet-1K in Table 7.

Table 5. Hyper-parameter settings for CIFAR-10/100.

Parameter	Value
#Iteration	1000
Optimizer	Adam with $\{\beta_1, \beta_2\} = \{0.5, 0.9\}$
Learning Rate	0.25 using cosine decay
Regularization Iterations	Last 500
γ	50
λ_1, λ_2	0.2, 4.0

Table 6. Hyper-parameter settings for Tiny-ImageNet.

Parameter	Value
#Iteration	2000
Optimizer	Adam with $\{\beta_1, \beta_2\} = \{0.5, 0.9\}$
Learning Rate	0.25 using cosine decay
Regularization Iterations	Last 1000
γ	50
λ_1, λ_2	0.2, 4.0

Table 7. Hyper-parameter settings for ImageNet-1K.

Parameter	Value
#Iteration	2000
Optimizer	Adam with $\{\beta_1, \beta_2\} = \{0.5, 0.9\}$
Learning Rate	0.25 using cosine decay
Regularization Iterations	Last 1000
γ	1.0
λ_1, λ_2	0.01, 0.25

9. Theoretical Analysis: Covariance Regularization for Dataset Distillation

Given a pre-trained model with layer-wise features $\boldsymbol{f}^l(\boldsymbol{x})$, our dataset distillation method enforces structural constraints on the covariance matrix of final-layer features for synthetic samples $s \in \mathcal{S}$:

$$\boldsymbol{Cov}^L_{\mathcal{S},c} = \frac{1}{|\mathcal{S}_c|}\left[(\boldsymbol{f}^L(\boldsymbol{s}) - \boldsymbol{\mu}^L_{\mathcal{S},c})(\boldsymbol{f}^L(\boldsymbol{s}) - \boldsymbol{\mu}^L_{\mathcal{S},c})^\top\right], \tag{10}$$

where $\boldsymbol{\mu}^L_{\mathcal{S},c}$ is the feature mean of the final layer. We propose the dual regularization objectives:

(1) Diagonal Maximization: Maximize variance of individual dimensions

$$\max \ \mathrm{Diag}(\boldsymbol{Cov}^L_{\mathcal{S},c}). \tag{11}$$

(2) Off-Diagonal Minimization: Decouple feature correlations

$$\min \ \|\boldsymbol{Cov}^L_{\mathcal{S},c} - \mathrm{Diag}(\boldsymbol{Cov}^L_{\mathcal{S},c})\|_F^2. \tag{12}$$

These objectives collectively enhance feature expressiveness while reducing redundant correlations.

Theorem 1 (Gradient Stability Guarantee)**.** *Let $h(\boldsymbol{s}) = \boldsymbol{W}\boldsymbol{f}^L(\boldsymbol{s}) + \boldsymbol{b}$ be the linear classifier for downstream fine-tuning. If the synthetic data covariance $\boldsymbol{Cov}^L_{\mathcal{S},c}$ satisfies:*

1. *$Var(\boldsymbol{f}^L_i) \geq \sigma^2_{\min}$ $(\forall i)$*
2. *$|Cov(\boldsymbol{f}^L_i, \boldsymbol{f}^L_j)| \leq \epsilon$ $(\forall i \neq j)$*

then the gradient matrix $\nabla_{\boldsymbol{W}}\mathcal{L}$ satisfies:

1. *Lower bounded Frobenius norm: $\|\nabla_{\boldsymbol{W}}\mathcal{L}\|_F \geq C\sigma_{\min}$*
2. *Upper bounded condition number: $\kappa(\nabla_{\boldsymbol{W}}\mathcal{L}) \leq \frac{\sigma^2_{\max}+(d-1)\epsilon}{\sigma^2_{\min}-(d-1)\epsilon}$*

Proof:

Step 1: Gradient Expression. For cross-entropy loss:

$$\nabla_{\boldsymbol{W}}\mathcal{L} = \mathbb{E}_{\mathcal{S},c}\left[(\boldsymbol{p} - \boldsymbol{y}) \otimes \boldsymbol{f}^L(\boldsymbol{s})\right] \tag{13}$$

where $\boldsymbol{p}$ represents the prediction vector, and $\otimes$ denotes the outer product.

Step 2: Variance Analysis. The Frobenius norm square of the gradient matrix reflects the diversity of update directions. The Frobenius norm can be calculated as follows:

$$\|\nabla_{\boldsymbol{W}}\mathcal{L}\|_F^2 = \mathrm{Tr}\left(\mathbb{E}[(\boldsymbol{p} - \boldsymbol{y})(\boldsymbol{p} - \boldsymbol{y})^\top]\mathbb{E}[\boldsymbol{f}^L(\boldsymbol{f}^L)^\top]\right), \tag{14}$$

Assuming that the prediction error $\boldsymbol{p} - \boldsymbol{y}$ is approximately independent of the feature $\boldsymbol{f}^L$, then:

$$\|\nabla_{\boldsymbol{W}}\mathcal{L}\|_F^2 \geq \lambda_{\min}(\boldsymbol{Cov}^L_{\mathcal{S}_c}) \cdot \mathbb{E}\|\boldsymbol{p} - \boldsymbol{y}\|^2. \tag{15}$$

According to the Gershgorin Circle Theorem, We can conclude that:

$$\lambda_{\min} \geq \sigma^2_{\min} - (d-1)\epsilon, \tag{16}$$

where d is the dimension of the matrix, and ϵ is the maximum absolute value of the non-diagonal elements of the covariance matrix. When ϵ is small enough, it satisfies $\lambda_{\min} \approx \sigma^2_{\min}$.

Step 3: Condition Number Analysis. The condition number of the covariance matrix is given by:

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} \leq \frac{\sigma_{\max}^2 + (d-1)\epsilon}{\sigma_{\min}^2 - (d-1)\epsilon}, \tag{17}$$

When $\epsilon \to 0$ and $\sigma_{\min}^2 \approx \sigma_{\max}^2$, we obtain $\kappa \approx 1$. A low condition number ensures the optimization landscape remains nearly isotropic, preventing gradient directions from being dominated by a small number of feature axes.

Step 4: Information Entropy Association. For a diagonally dominant covariance matrix with eigenvalue distribution approximating independent Gaussians, the differential entropy is:

$$H(\boldsymbol{f}^L) = \frac{1}{2}\ln\left[(2\pi e)^d \det(\boldsymbol{Cov}_{\mathcal{S}})\right] \geq \frac{1}{2}\ln(2\pi e\sigma_i^2) \tag{18}$$

The entropy $\sum \sigma_i^2$ is maximized under fixed trace if and only if $\sigma_i = \sigma_j$. Our proposed constraints enhance entropy by guaranteeing minimum variance across dimensions, thereby strengthening feature representation.

10. Analysis of Computational Overhead

We report the runtime and GPU memory usage on ImageNet-1K in Table 8. Compared to LPLD, our DEDA reduces training time by 3.6 hours and GPU memory consumption by 0.89 GB on a single A100 GPU, as it aligns the mean and covariance of pooled features, whereas LPLD operates on the unpooled high-dimensional features.

Table 8. Computational cost on Imagenet-1K with IPC=50.

Method	Runtime	Memory
LPLD	123s×1000 = 34.16h	2.79GB
Ours	110s×1000 = 30.56h	1.90GB

11. Feature Cosine Distance Calculation

As shown in Figure 5, we compute the average cosine distance between all distilled samples and their corresponding class prototypes for the first 20 classes of CIFAR-100 under IPC=50. Our method achieves a 0.0223 improvement over baseline approaches. This quantitative gain demonstrates that DEDA effectively enhances the semantic diversity of distilled data, as reflected by increased intra-class feature dispersion in the embedding space. This improvement arises from the limitations of SRe2L, where aligning all class-distilled data at the same BN layers restricts diversity. By using Gaussian distribution matching and introducing covariance regularization in the last layer, we better preserve feature diversity and explicitly increase the semantic spread within each class.

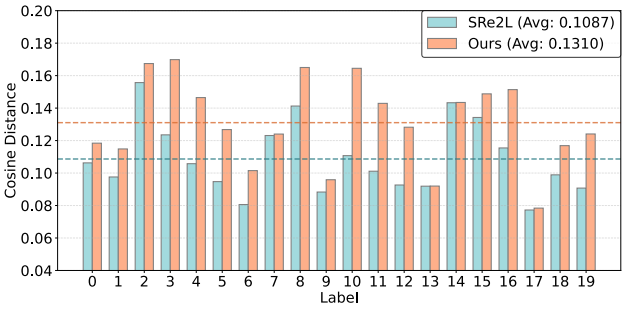

Figure 5. Cosine distances between samples and class centers for the first 20 Classes in CIFAR-100 distilled data (IPC=50).

12. Visualization of More Distilled Data

More visualization results of distilled data randomly sampled from DEDA are shown in Figure 6 (CIFAR-10), Figure 7 (CIFAR-100), and Figure 8 (ImageNet-1K).

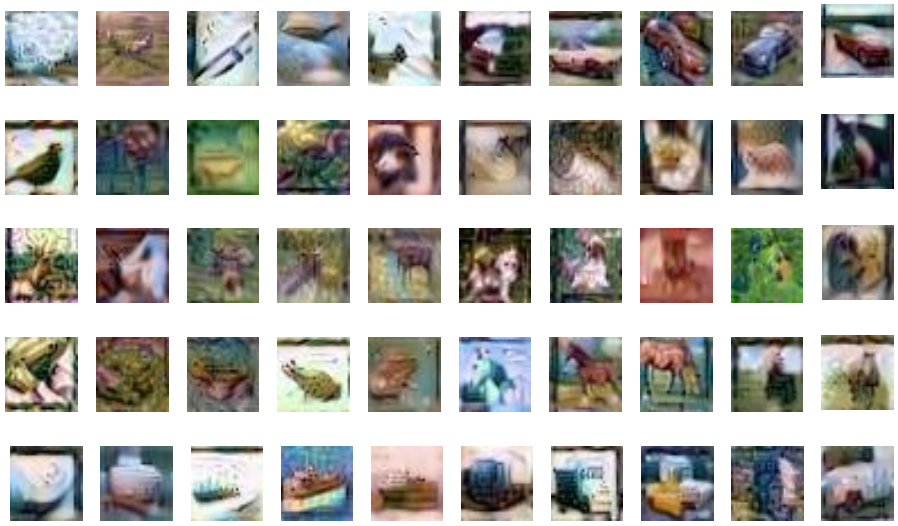

Figure 6. Visualization of distilled data generated by DEDA on CIFAR-10.

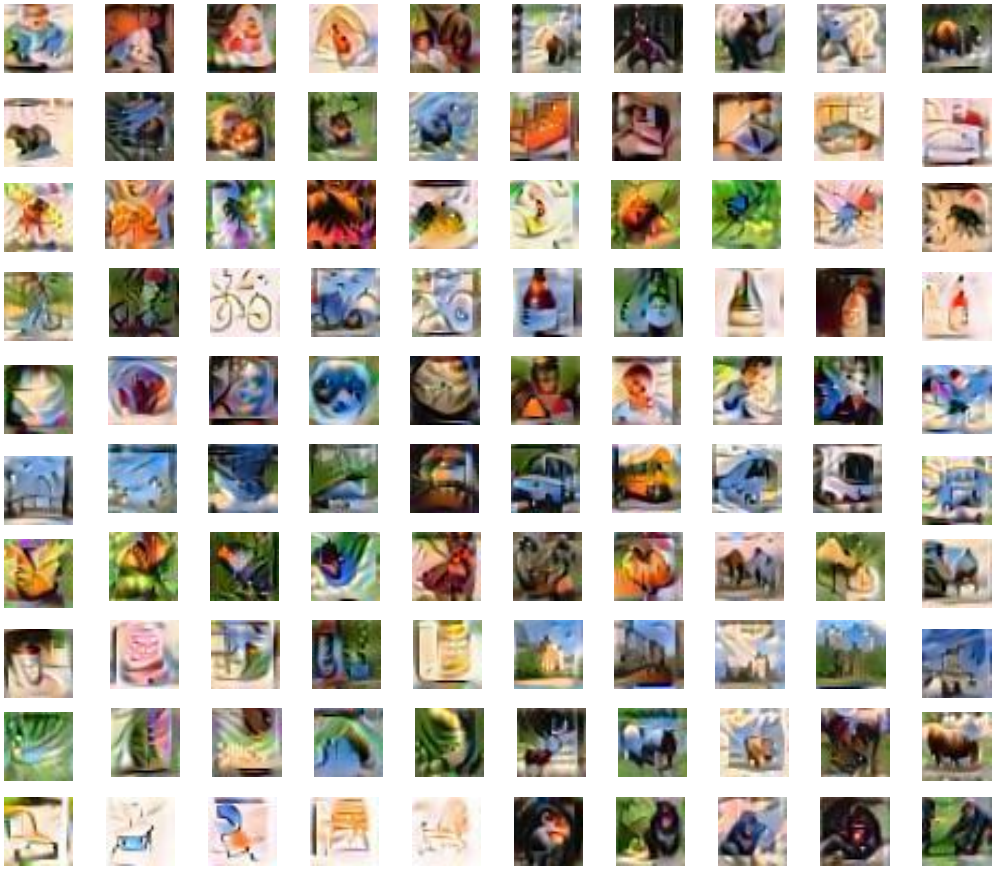

Figure 7. Visualization of distilled data generated by DEDA on CIFAR-100.

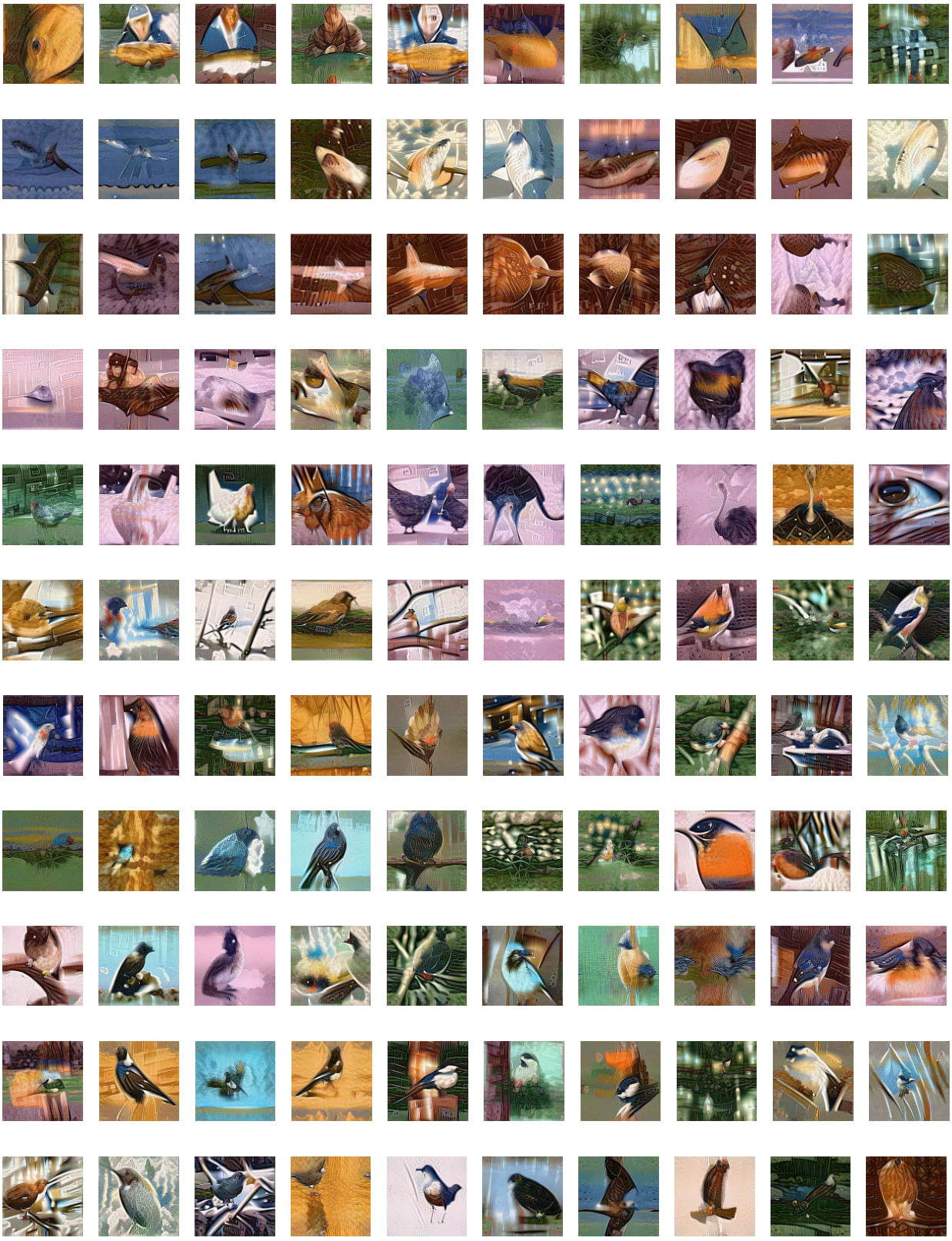

Figure 8. Visualization of distilled data generated by DEDA on ImageNet-1K.