Deep Noisy-Label Learning via Effective Rank Reduction

Changyi Ma[1*], Zihan Fang[1,2*], Lihua Zhou[3], Tao Li[4], Wenyu Liu[2†], Runsheng Yu[5†], and Xuan Song[1]

[1] School of Artificial Intelligence, Jilin University, China
[2] College of Computer and Cyber Security, Fujian Normal University, China
[3] Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong, China
[4] Institute of Image Processing and Pattern Recognition, School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, China
[5] China Merchants Research Institute of Advanced Technology, China Merchants Group, Hong Kong, China

[*] Equal contribution.
[†] Corresponding authors. Emails: liuwenyu@fjnu.edu.cn; runshengyu@gmail.com

Abstract. Learning with noisy labels in large-scale classification remains challenging. Existing forward-correction methods estimate class-dependent label noise via a transition matrix, but this matrix is typically assumed to be full-rank and unconstrained. However, real-world label noise is typically structured and exhibits low-rank properties due to semantic clustering, especially when the number of classes is large. This mismatch between modeling assumptions and the intrinsic structure of label noise leads to unstable optimization and degraded generalization performance. Motivated by this observation, we propose a novel **L**ow-**E**ffective-Rank **N**oisy-**L**abel Learning (**LENL**) framework, which explicitly constrains the transition matrix via nuclear-norm regularization. This low-rank inductive bias captures structured confusion patterns while suppressing spurious noise modes, yielding both statistical efficiency and improved generalization. We further provide theoretical analysis establishing the benefits of rank-constrained estimation under structured noise assumptions. To enable stable and scalable optimization, we introduce a Newton–Schulz optimizer tailored for label-noise learning, which approximates the nuclear norm through decomposition-free iterative updates, avoiding the numerical instability and computational overhead of explicit singular value decomposition. Extensive experiments on CIFAR-100, CIFAR-100N, ImageNet-1K, and Noisy Ostracods demonstrate consistent and significant improvements over state-of-the-art label-noise learning methods.

Keywords: Noisy Label Learning · Effective Rank

1 Introduction

Real-world datasets are rarely perfectly annotated because human errors, ambiguous visual content, and automated labeling pipelines inevitably introduce label noise. Deep networks can easily memorize corrupted annotations given their high capacity, leading to severe generalization degradation [43]. Consequently, robust learning with noisy labels has become a central problem in large-scale classification.

Existing approaches can be broadly categorized into *transition matrix estimation* and *sample-selection* methods. In this paper, we focus on the former, particularly forward-correction–based frameworks [17, 23, 39], which explicitly model class-dependent noise via a transition matrix T. These methods typically treat T as a full-rank matrix and estimate all C^2 entries independently.

However, this full-rank modeling strategy makes no structural assumptions about class confusion and may therefore overparameterize noise patterns. In practice, real-world label confusion is rarely arbitrary. Annotators tend to confuse semantically or visually similar classes (e.g., *oak tree*, *pine tree*, and *willow tree*), resulting in correlated or nearly collinear rows in T. When classes naturally cluster into a smaller number of semantic groups, the spectrum of T decays rapidly, indicating reduced effective rank, a phenomenon we empirically verify in Section 3.1. In such settings, full-rank estimation allocates capacity to tail modes that capture unstructured corruption rather than systematic confusion, leading to statistical inefficiency and degraded performance.

Motivated by this observation, we propose *Low-Effective-Rank Noisy-Label Learning* (*LENL*), a unified framework that explicitly exploits the reduced effective-rank structure. Concretely, LENL incorporates nuclear-norm regularization—the tightest convex relaxation of matrix rank [24]—into the forward-correction objective. Specifically, LENL jointly *i)* fits noisy observations using a forward-corrected cross-entropy loss, and *ii)* constrains the effective rank of T via nuclear-norm regularization, encouraging structured and compact confusion modeling.

A key technical challenge arises from optimizing the nuclear norm. Standard differentiable SVD-based implementations [11] become numerically unstable when T exhibits many near-zero singular values, precisely the regime induced by effective-rank constraints. To address this issue, we develop a **Newton–Schulz (NS)** [27] optimizer, a decomposition-free iterative method that computes the nuclear norm using matrix multiplications only, ensuring numerical stability and scalability.

We evaluate LENL on both *synthetic* and *real-world* noisy benchmarks, including CIFAR-100, CIFAR-100N, ImageNet-1K, Noisy Ostracods, and Food-101N. Across symmetric and pair-flip synthetic noise as well as naturally noisy datasets, LENL consistently outperforms state-of-the-art transition-based methods.

In summary, our contributions are:

- **Reduced Effective Rank Structure in Label Noise.** We empirically and theoretically demonstrate that noise transition matrices in both real-world

and synthetic benchmarks exhibit reduced effective rank due to semantic clustering and feature overlap.
- **Structured Low-Rank Transition Modeling.** We introduce LENL, the first framework that explicitly incorporates effective rank constraints into transition-based noisy-label learning, along with theoretical generalization guarantees under structured noise assumptions. We further propose a numerically stable Newton–Schulz optimizer for efficient nuclear-norm regularization.
- **Strong and Consistent Empirical Results.** We show consistent improvements over state-of-the-art methods on CIFAR-100, ImageNet-1K, and real-world noisy datasets, with ablations validating each component.

2 Preliminary

2.1 Problem Setup

Let $(X, Y) \in \mathcal{X} \times \{1, \ldots, C\}$ denote the random variables for instances and clean labels, where $\mathcal{X}$ is the instance space and C is the number of classes. In many real-world applications [29], however, clean labels are unobservable and only noisy labels are available. Let $\tilde{Y}$ denote the random variable corresponding to noisy labels. We are given a training sample $\{(x_1, \tilde{y}_1), \ldots, (x_n, \tilde{y}_n)\}$ drawn from the noisy distribution $\mathcal{D}_{\text{noisy}}$ over $(X, \tilde{Y})$. Our goal is to learn a robust classifier that predicts the underlying clean labels for unseen test instances using only noisy training data.

A mainstream approach for robust learning under label noise introduces a transition matrix to model the relationship between the clean labels Y and the noisy labels $\tilde{Y}$ [2, 3]. This formulation enables the recovery of clean class posteriors from noisy observations.

Formally, let $P(Y|x) = [P(Y = 1|x), \ldots, P(Y = C|x)]^\top$ denote the clean class posterior vector, and $P(\tilde{Y}|x) = [P(\tilde{Y} = 1|x), \ldots, P(\tilde{Y} = C|x)]^\top$ denote the noisy class posterior vector. These distributions are linked by an instance-independent noise transition matrix $T \in \mathbb{R}^{C \times C}$, where the entry $T_{ij} := P(\tilde{Y} = j \mid Y = i)$ denotes the probability of a ground-truth label i being misclassified as label j. Under the commonly adopted class-conditional noise assumption, the noisy posterior satisfies:

$$P(\tilde{Y} = j|x) = \sum\nolimits_{i=1}^{C} P(\tilde{Y} = j|Y = i, x)P(Y = i|x) = \sum\nolimits_{i=1}^{C} T_{ij}P(Y = i|x) \tag{1}$$

Throughout this paper, the transition matrix T is defined at the *population level*, characterizing how *clean* labels are corrupted into *noisy* ones. In contrast, the classifier $p_\theta(x)$ estimates $P(Y|x)$ at the *sample level* and is the only quantity directly optimized during training.

2.2 Related Work

Transition Matrix Estimation Methods. Transition matrix-based methods include forward correction (multiplying model outputs by $T^\top$) and backward cor-

rection (multiplying labels by T) [23]. Backward correction requires transition matrix inversion, causing severe numerical instability when T is ill-conditioned. Forward correction [23] pioneered the use of transition matrices T to model label noise, requiring a small clean set for estimation. Subsequent works eliminated this requirement through anchor points [20] or structural assumptions: VolMinNet [17] uses volume minimization with provable guarantees, Dual-T [41] separates instance- and class-dependent noise, and Class2Simi [36] exploits class similarity. PLM [45] takes a different approach by generating part-level labels through instance cropping and introducing a single-to-multiple transition matrix to model the relationship between noisy and part-level labels. Despite these advances, existing transition-based methods generally treat the transition matrix as fully expressive and do not explicitly exploit reduced effective-rank structure, which commonly arises in semantically structured natural datasets. Moreover, numerical stability becomes increasingly critical in such regimes.

Sample Selection Methods. Sample selection methods rely on the memorization effect of deep networks to filter out noisy examples. Co-teaching [7] trains two networks that iteratively select clean samples for each other. DivideMix [16] combines Gaussian Mixture Models with semi-supervised learning, while MOIT [22], Jo-SRC [40], and UNICON [13] add joint training and contrastive objectives. Although these approaches avoid explicit estimation of T, they typically lack theoretical guarantees on the clean (noise-free) risk. In contrast, our work provides provable generalization bounds (Theorem 1) that explicitly control the clean loss when training solely on noisy labels.

Multi-Component Methods. Recent approaches combine multiple techniques such as sample selection, contrastive learning, and data augmentation (e.g., DivideMix [16], UNICON [13]) to achieve strong empirical performance. While effective, these methods integrate several orthogonal components, making it difficult to isolate the contribution of any single mechanism. In contrast, our work focuses on a single principled component—reduced-rank regularization of the transition matrix—providing clearer attribution of performance gains.

Other Methods for Noisy-Label Learning. Other approaches include label correction [4, 10], regularization methods [19, 30, 34], robust loss functions [21,33], meta-learning methods [1,44], and training-based adaptation methods [37, 42]. Since our focus is on transition matrix estimation, these methods are orthogonal to our work.

2.3 Effective Rank

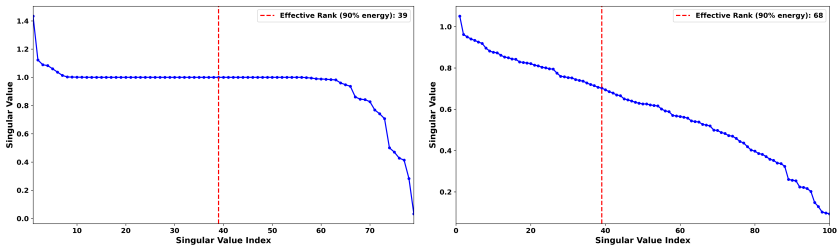
We interpret the singular values of T as an energy distribution. Let the singular value decomposition of T be $T = U\Sigma V^\top$, where the singular values satisfy $\sigma_1 \geq \cdots \geq \sigma_C \geq 0$. Following [31], we define the proportion of spectral energy captured by the top-r components as:

$$E_r(T) := \frac{\sum_{i=1}^{r} \sigma_i^2}{\sum_{i=1}^{C} \sigma_i^2} \tag{2}$$

The **effective rank** r_{eff} is defined as the smallest r such that $E_r(T) \geq 0.90$ [12]. This notion captures the phenomenon that, although T may be full rank, most of its spectral energy can be concentrated in $r < C$ dominant modes due to rapid singular value decay.

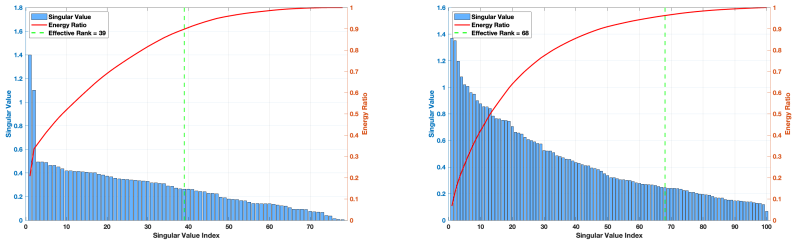
Compatibility with Prior Work. Prior works [17, 23] assume that the transition matrix is full rank to ensure invertibility. This assumption does not contradict our reduced effective-rank perspective. Full rank and low effective rank can coexist: a matrix may have all singular values strictly positive (hence full rank), while the majority of its spectral energy is concentrated in a small subset of dominant components.

3 Methodology



(a) Noisy Ostracods **(b)** CIFAR-100N

Fig. 1: Singular Values vary on various datasets.



(a) Noisy Ostracods **(b)** CIFAR-100N

Fig. 2: Singular Values and Energy Ratio vary with Singular Value Index on various datasets.

3.1 Reduced Effective Rank Observation

Row collinearity from semantic confusion. Semantically similar classes exhibit nearly identical confusion patterns due to **feature overlap**: when classes share visual or conceptual attributes, annotators make systematic mistakes across

these classes [35,38]. For example, *oak tree*, *pine tree*, and *willow tree* in CIFAR-100N (all from the superclass *Trees*) are confused at similar rates, resulting in approximately collinear rows in T. We verify this using cosine similarity $s_{ij} := \cos(T_{i,:}, T_{j,:})$ [28].

Figure 3 shows that all tree species pairs in CIFAR-100N exhibit $s_{ij} > 0.75$, confirming nearly identical confusion patterns.

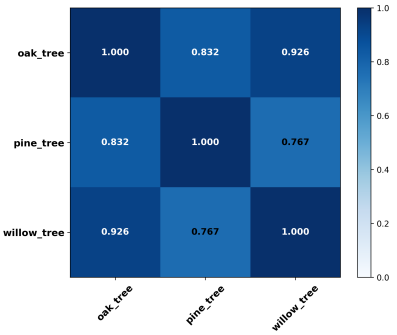


Fig. 3: Illustrative Example of the Colinearity within the noisy label transition matrix. Each value represents the similarity scores among these three classes.

Singular value decay. This high row collinearity directly induces reduced effective rank: when multiple rows of T are approximately collinear, the matrix can be well-approximated by a low-dimensional subspace, causing its singular values to decay rapidly. Real-world data matrices with such hierarchical structure are known to exhibit rapid singular value decay [31]. Figure 2 shows fast singular value decay in learned transition matrices, with effective ranks $r_{\text{eff}} = 30$ and 68 (green lines) capturing 70–90% of spectral energy (red curves). Figure 1 shows a sudden drop-off in singular values.

Effective rank concentration. We then verify the reduced effective rank property of transition matrices. Table 1 shows that natural noisy datasets achieve $r_{\text{eff}}/C \in [0.68, 0.75]$: CIFAR-100N achieves $r_{\text{eff}}/C = 0.68$, while Noisy Ostracods achieves 0.75. These effective-rank ratios are comparable to or lower than synthetic pair noise (0.78), but substantially lower than symmetric noise (0.86). This confirms our hypothesis that human annotators systematically confuse semantically similar classes, producing an exploitable reduced effective rank structure. Moreover, structured synthetic noise also exhibits moderate rank concentration, validating that our regularization approach generalizes beyond real-world data.

Reduced-rank regularization necessity. Full-rank methods that estimate all C^2 entries of T waste model capacity: they dedicate equal parameters to both the signal-carrying top-r modes (systematic confusion patterns) and the tail $(C - r)$ modes (noise), leading to overfitting and degraded generalization. Ablation studies (Section **??**) further confirm that removing reduced effective rank regularization significantly degrades performance (CIFAR-100N: -4.6% accuracy), validating the necessity of explicit rank constraints.

These observations consistently indicate that real-world transition matrices possess a pronounced reduced effective-rank structure. Preserving this structural property is therefore not merely descriptive, but potentially beneficial for improving generalization under noisy supervision. Consequently, we are motivated to *explicitly* impose a reduced effective rank constraint on T via nuclear

norm regularization, focusing learning on the signal-dominant subspace while suppressing noise-dominated modes.

Table 1: Effective Rank (eRank) of various transition matrices T on synthetic noisy datasets (CIFAR-100, ImageNet-1K) and natural noisy datasets (CIFAR-100N, Noisy Ostracods).

	Synthetic Noisy Datasets				Natural Noisy Datasets	
	CIFAR-100		ImageNet-1K		CIFAR-100N	Noisy Ostracods
	Pair-20%	Pair-45%	Pair-20%	Pair-45%		
Row Size	100	100	1000	1000	100	78
eRank	86	69	857	686	68	39

3.2 Proposed Method

Motivated by the reduced effective-rank observation, we propose ***L**ow-**E**ffective-**R**ank **N**oisy-**L**abel **L**earning*, a unified framework that jointly learns the classifier and a rank-regularized transition matrix. Concretely, we optimize the following objective:

$$\mathcal{L}(\theta, T) := \underbrace{\mathbb{E}_{(x,\tilde{y})\sim\mathcal{D}}\Big[-\tilde{y}^\top \log\big(T^\top p_\theta(x)\big)\Big]}_{\text{Forward-corrected CE}} + \underbrace{\lambda\|T\|_*}_{\text{Regularizer}} \tag{3}$$

where θ denotes the network parameters, $p_\theta(x) \in \Delta^{C-1}$ is the softmax posterior over clean classes, $\tilde{y} \in \{0,1\}^C$ is the observed noisy one-hot label, and $T \in \mathbb{R}^{C\times C}$ is the row-stochastic noise transition matrix. The first term is the forward-corrected cross-entropy loss, which fits the noisy-label distribution through $T^\top p_\theta(x)$. The second term controls the spectral complexity of T through nuclear-norm regularization, with $\lambda > 0$ balancing noisy-label fitting and spectral regularization.

The nuclear norm serves as a tractable surrogate for rank-like spectral complexity. Directly optimizing the effective rank is difficult because it is a discrete and non-smooth function of the singular spectrum. Therefore, instead of imposing a hard effective-rank constraint, we use the nuclear norm as a soft convex surrogate that encourages the singular spectrum of T to concentrate on dominant modes while suppressing non-essential tail modes.

Although T and p_θ are jointly optimized in Eq. (3), degenerate factorizations are discouraged by the structure of the objective. First, T is constrained to be row-stochastic, which removes scaling ambiguity between the transition matrix and the classifier output. Second, the cross-entropy loss is applied to the forward-corrected prediction $T^\top p_\theta(x)$. If T collapses to a rank-one matrix, or to a matrix with nearly identical rows, then $T^\top p_\theta(x)$ becomes restricted to a low-dimensional convex hull and is nearly insensitive to the input-dependent classifier output. Such a collapsed transition matrix cannot adequately account for diverse, noisy multi-class labels and would therefore be penalized by the cross-entropy term.

Theoretical Support Beyond empirical motivation, reduced effective rank also plays a role in controlling model complexity. Since the transition matrix is jointly learned with the classifier, constraining its spectral complexity can reduce hypothesis capacity and tighten generalization guarantees. We next demonstrate how the objective (3) can be derived to tighten the generalization bound.

Theorem 1 (Generalization bound). *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to \Delta^{C-1}\}$, and let $T \in \mathbb{R}^{C\times C}$ be a row-stochastic noise-transition matrix. Assume that for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$, the forward-corrected predictions satisfy $(T^\top f(x))_j \geq c > 0, \quad \forall j \in \{1,\dots,C\}$, where c is a constant. Given n i.i.d. noisy samples $D = \{(x_i, \tilde{y}_i)\}_{i=1}^n$ drawn from the noisy distribution $\mathcal{D}_{noisy}$, define the forward-corrected empirical risk $\widehat{R}_n^{\text{forw}}(f) := -\frac{1}{n}\sum_{i=1}^n \tilde{y}_i^\top \log\big(T^\top f(x_i)\big)$, and the population noisy risk $R_{noisy}(f) := -\mathbb{E}_{(x,\tilde{y})\sim\mathcal{D}_{noisy}}[\tilde{y}^\top \log(T^\top f(x))]$. Then, with probability at least $1-\delta$ over the draw of D,*

$$R_{noisy}(f) \leq \widehat{R}_n^{\text{forw}}(f) + \frac{2}{c}\|T\|_*\widehat{\mathfrak{R}}_n(\mathcal{F}) + M\sqrt{\frac{\log(2/\delta)}{2n}},$$

where $\|T\|_*$ is the nuclear norm of T, $\widehat{\mathfrak{R}}_n(\mathcal{F})$ is the empirical Rademacher complexity of $\mathcal{F}$ on D, and $M = \log(C/c)$.

Moreover, when $\sigma_{\min}(T) > 0$, we have

$$\min_f R_{clean}(f) \leq \min_f \widehat{R}_n^{\text{forw}}(f) + 2\frac{1}{c}\|T\|_*\widehat{\mathfrak{R}}_n(\mathcal{F}) + M\sqrt{\frac{\log(2/\delta)}{2n}},$$

where $R_{clean}(f) := -\mathbb{E}_{(x,y)\sim\mathcal{D}_{clean}}[y^\top \log f(x)]$.

The theorem shows that minimizing an upper bound of the noisy-label generalization error naturally leads to the objective in Eq. (3), thereby providing theoretical justification for nuclear-norm regularization as a principled structural prior on T.

3.3 Stable LENL Algorithm

However, computing the nuclear norm $\|T\|_* = \sum_i \sigma_i(T)$ requires singular value decomposition, which introduces both computational and numerical challenges in practice:

Numerical instability. When T has reduced effective rank (e.g., Table 1), many singular values cluster near zero. Computing $\|T\|_*$ requires SVD. The gradient of the nuclear norm involves $(\sigma_i^2 - \sigma_j^2)^{-1}$ terms that become ill-conditioned in this regime, causing training instability. This issue is widely observed in automatic differentiation frameworks; for instance, PyTorch's `torch.svd` is known to produce NaN gradients for reduced-rank matrices.

To retain the benefits of reduced-effective-rank regularization while ensuring numerical stability, we replace the exact nuclear norm with a matrix-function surrogate that can be evaluated through fast, decomposition-free iterations.

Specifically, we approximate the matrix square root via the Newton–Schulz (NS) iteration [27], a quadratically convergent, multiplication-based scheme that

is free of explicit decompositions. Starting from $Y_0 = \alpha(TT^\top + \epsilon I)$ and $Z_0 = \alpha I$, where $\alpha = 1/\|TT^\top + \epsilon I\|_F$ ($\|\cdot\|_F$ is the Frobenius norm)), the coupled updates:

$$Y_{k+1} = \tfrac{1}{2} Y_k(3I - Z_k Y_k), \quad Z_{k+1} = \tfrac{1}{2}(3I - Z_k Y_k) Z_k$$

converge to $(TT^\top + \epsilon I)^{1/2}$ quadratically when $\|I - Y_0 Z_0\| < 1$, which is ensured by the scaling α. We set $\epsilon = 10^{-6}$ and use $k = 5$ iterations in our experiments.

Overall, the NS surrogate eliminates eigenvalue computation, thereby providing numerical stability for T.

Overview We denote the variant using exact SVD for nuclear-norm gradient computation as *LENL-SVD*, and the variant using the Newton–Schulz approximation as *LENL-NS*.

Algorithm **??** (in Appendix **??**) summarizes the scalable LENL procedure. Each mini-batch is processed in three steps:

First, the network's class probabilities are multiplied by the transition matrix, and cross-entropy against noisy labels is computed, keeping the optimization aligned with the unknown clean distribution.

Second, to encourage reduced effective rank, the algorithm computes the nuclear norm regularizer either exactly via SVD or efficiently through Newton–Schulz iterations, thereby avoiding numerical instability.

Finally, the combined loss drives joint SGD updates of both classifier and transition matrix, with row-wise softmax projection maintaining row stochasticity.

4 Evaluation

4.1 Evaluation Setup

Datasets We evaluate our method on five benchmark datasets: two synthetic-noise datasets (CIFAR-100 [14] and ImageNet-1K [26]) and three real-world noisy datasets (CIFAR-100N [35], Noisy Ostracods [9], and Food-101N [15]), as summarized in Table 2. These datasets are widely adopted in prior noisy-label learning studies [6, 7, 23, 25]. Table 2 provides detailed information for each dataset.

Since *CIFAR-100* [14] and *ImageNet-1K* [26] are clean datasets, we follow the standard protocol [7, 23, 25] and corrupt them synthetically using a noise transition matrix T, where $T_{ij} = P(\tilde{y} = j | y = i)$ denotes the probability that clean label $y = i$ is flipped to noisy label $\tilde{y} = j$.

We consider two representative noise models [7, 32]:

i) Symmetric flipping: A uniform noise model where any class can be confused with any other class with equal probability. Formally, $T_{ij} = \epsilon/(C-1)$ for $i \neq j$ and $T_{ii} = 1 - \epsilon$, where ϵ denotes the noise ratio.

ii) Pair flipping: A structured noise model simulating fine-grained classification scenarios where confusion occurs only between specific, visually or semantically similar class pairs (e.g., dog breeds). Each class i is assigned a designated confusing class j, with $T_{ij} = \epsilon$ and $T_{ii} = 1 - \epsilon$.

Table 2: Summary of dataset information.

Datasets	# of training	# of testing	# of classes
CIFAR-100 [14]	50,000	10,000	100
ImageNet-1K [26]	1,281,167	100,000	1,000
CIFAR-100N [35]	50,000	10,000	100
Noisy Ostracods [9]	66,352	6,896	52
Food-101N [15]	310,000	25,000	101

Following standard benchmarks [7,45], we evaluate on noise rates $\epsilon \in \{0.2, 0.5\}$ for symmetric flipping and $\epsilon \in \{0.2, 0.45\}$ for pair flipping.

Baseline Methods We compare our proposed LENL (Algorithm **??**) with the following state-of-the-art methods: CE (vanilla training), VolMinNet (forward transition matrix correction), Co-teaching (sample selection), and PLM (one of the latest works).

i) CE [5]: cross-entropy (CE) serves as the baseline method.

ii) VolMinNet [17]: it estimates the transition matrix by incorporating a minimum-volume constraint for noise modeling.

iii) Co-teaching [7]: a sample-selection method that trains two networks and exchanges small-loss samples for mutual learning, representing the alternative paradigm to transition-matrix approaches.

iv) PLM [45]: it generates multiple part-level labels to guide training, representing the strongest overall baseline in our comparison.

v) IDLE [18]: this method enhances DCNN robustness against label noise by approximating instance-dependent noise transition matrices through inter-class correlation within mini-batches.

Implementation Details We implement two variants of our method in different environments:

i) LENL-SVD, which computes the exact nuclear norm and its gradients via SVD. We use it for relatively small-class datasets (e.g., CIFAR-100 with $C = 100$), where the SVD remains numerically stable. We set $\lambda = 5 \times 10^{-4}$ for all datasets. ***ii) LENL-NS***, which employs the Newton-Schulz iteration for efficient and stable approximation. This is designed for large-class datasets (e.g., ImageNet-1K with $C = 1000$) and real-world noisy datasets, where reduced effective rank leads to many near-zero eigenvalues, causing numerical instability in SVD, as discussed in Section 3.3. We set $\lambda = 5 \times 10^{-4}$ for all datasets.

All methods are implemented in PyTorch with the same backbone architecture and trained on a single NVIDIA V100 GPU for fair comparison. We report baseline results from original publications, when available on the same datasets; otherwise, we implement them using official code or following published hyperparameters.

i) **CIFAR-100**. We use a ResNet-32 [8] as a backbone network. We use SGD to train the network and the transition matrix, with an initial learning rate of 0.05, weight decay of 0.001, and momentum of 0.9. The learning rate is divided by 100 after the 40th epoch. We apply random horizontal flips and random crops of size 32×32 pixels after adding a 4-pixel padding on each side for the CIFAR-100 datasets. The results for PLM, VolMinNet, and Co-Teaching are taken from [45]. The results for CE on Pair-45% and Sym-50% are from [7]. We implement CE on Pair-20% using the official code from [7], as this setting was not reported in their paper.
ii) **ImageNet-1K**. Following [8], we use a ResNet-50 as a backbone network. SGD optimizer with a mini-batch size of 256 is used to train the network, with an initial learning rate of 0.1, weight decay of 0.0001, and momentum of 0.9. For our method, the learning rate is divided by 10 after the 30th epoch and every 15 epochs thereafter. We resize the images to 256 × 256 and apply random horizontal flips and random crops of size 224 × 224 pixels for the ImageNet dataset. All baseline methods are implemented using their official code.
iii) **CIFAR-100N**. Following [35], we use a ResNet-32 [8] as a backbone network. SGD optimizer with a mini-batch size of 128 is used to train the network and the transition matrix, with an initial learning rate of 0.05, weight decay of 0.001, and momentum of 0.9. The learning rate is divided by 100 after the 40th epoch. We apply random horizontal flips and random crops of size 32×32 pixels after adding a 4-pixel padding on each side for the CIFAR-100N dataset.
iv) **Noisy Ostracods**. Following [9], we use a pre-trained ResNet-50 as a backbone network. SGD optimizer is used to train the network, with an initial learning rate of 0.1, weight decay of 0.0001, and momentum of 0.9. The learning rate is divided by 3.33 for every 15 epochs. The results for PLM and Co-Teaching are taken from [9]. VolMinNet is implemented using the official code, as this baseline was not reported in their paper.
v) **Food-101N**. We follow the CleanNet [15] setup, using a ResNet-50 pre-trained on ImageNet-1K as the backbone and fine-tuning it. An SGD optimizer is used to train the network with an initial learning rate of 0.001, weight decay of 0.0001, and momentum of 0.9. We train the model for 30 epochs, and the learning rate is divided by 3.33 every 10 epochs. We resize the images to 256 × 256 and apply random horizontal flips and random crops of size 224 × 224 pixels for Food-101N images.

Evaluation Metrics We employ test accuracy as the primary evaluation metric. $\boldsymbol{Accuracy} = \frac{\#\text{ of correct predictions}}{\#\text{ of test samples}}$ measures overall prediction correctness. High accuracy indicates good performance.

4.2 Performance Analysis

Performance on Synthetic Noisy Datasets Performance on CIFAR-100 Dataset. We show our results on the CIFAR-100 dataset. Table 3 shows the test accuracy under all four synthetic-noise settings on the CIFAR-100 dataset.

Our proposed LENL variants consistently outperform the strongest competitor by clear margins ($\geq 0.3\%$ on average). Notably, LENL-NS pushes absolute accuracy to 70.42% (Sym-20%) and 62.88% (Pair-45%).

Table 3: Comparison of accuracy and standard deviation (expressed as percentages) across five trials on the CIFAR-100 dataset under various synthetic noisy-label settings. The best classification accuracy is indicated in bold.

Methods/Datasets	**CIFAR-100**		**CIFAR-100**	
	Sym-20%	**Sym-50%**	**Pair-20%**	**Pair-45%**
CE [5]	47.55 ± 0.47	25.21 ± 0.64	50.43 ± 0.51	31.99 ± 0.64
VolMinNet [17]	64.94 ± 0.40	53.89 ± 1.26	68.45 ± 0.69	58.90 ± 0.89
Co-Teaching [7]	44.89 ± 0.63	32.19 ± 0.90	45.78 ± 1.02	28.98 ± 0.42
PLM [45]	69.54 ± 0.23	60.44 ± 0.28	72.67 ± 0.32	61.90 ± 1.82
ILDE [18]	70.08 ± 0.31	60.95 ± 0.42	72.68 ± 0.24	61.45 ± 0.56
LENL-SVD	**70.42 ± 0.18**	**61.35 ± 0.21**	**73.21 ± 0.15**	**62.88 ± 0.35**

Performance on ImageNet-1K Dataset. Table 4 shows the test accuracy under all four synthetic-noise settings on the ImageNet-1K dataset.

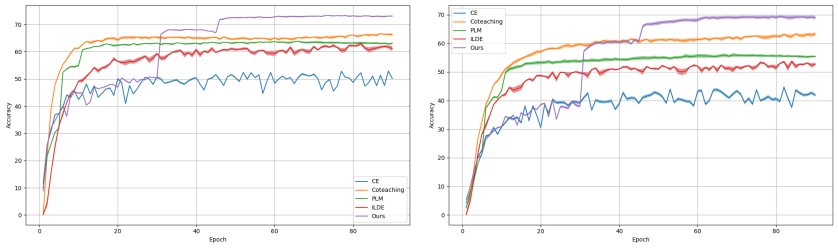
Our proposed LENL-NS method demonstrates robust and superior performance on the ImageNet-1K dataset under various synthetic noisy label settings. It achieves the highest classification accuracy across all noise types and ratios, with 73.35% under Sym-20%, 69.37% under Sym-50%, 72.59% under Pair-20%, and 65.59% under Pair-45%, substantially outperforming all baseline methods. These results validate the effectiveness of LENL-NS in handling diverse and challenging scenarios involving label noise.

Table 4: Comparison of accuracy and standard deviation (expressed as percentages) across five trials on the ImageNet-1K dataset under various synthetic noisy-label settings. The best classification accuracy is indicated in bold. '-' means that the method fails to complete due to numerical instability.

Methods/Datasets	ImageNet-1K		ImageNet-1K	
	Sym-20%	Sym-50%	Pair-20%	Pair-45%
CE [5]	52.93 ± 0.22	44.65 ± 0.66	51.98 ± 0.29	34.59 ± 0.63
VolMinNet [17]	-	-	-	-
Co-Teaching [7]	66.69 ± 0.34	63.40 ± 0.50	67.01 ± 1.28	64.08 ± 0.57
PLM [45]	63.87 ± 0.16	55.96 ± 0.74	64.22 ± 0.29	53.01 ± 0.43
ILDE [18]	62.86 ± 0.45	53.71 ± 0.40	62.81 ± 0.23	60.35 ± 0.52
LENL-NS (Ours)	**73.35 ± 0.08**	**69.37 ± 0.56**	**72.59 ± 0.21**	**65.59 ± 0.89**

Notably, VolMinNet encounters scalability challenges on large-class datasets such as ImageNet-1K ($C = 1,000$), due to its reliance on exact SVD of the transition matrix T, which becomes numerically unstable when T contains near-zero eigenvalues (see Section 3).

Convergence. To further verify the scalability of our LENL, we show the test accuracy vs. the number of epochs. Figure 4 shows the results. As can be seen, our method converges faster and to a higher final accuracy than the baselines.



(a) ImageNet-1K, Sysmetry 20% **(b)** ImageNet-1K, Sysmetry 50%

Fig. 4: Results on ImageNet 1K dataset. Testing accuracy varies with epochs. VolMinNet is omitted from the figure due to numerical instability.

Performance on Real-World Noisy Datasets We report accuracy for CIFAR-100N, Food-101N, and Noisy Ostracods datasets. For Noisy Ostracods, we additionally report macro-averaged precision, recall, and F1 score to evaluate the classification performance, following [9].

Performance on Noisy Ostracods Dataset. LENL-NS demonstrates superior or highly competitive performance across key metrics on the Noisy Ostracods dataset. LENL-NS achieves the highest scores in Accuracy (97.20%), Recall (77.93%), and F1-score (79.75%), while maintaining a strong Precision of 87.21%, outperforming other methods such as Co-Teaching, PLM, and CE. Notably, VolMinNet fails to complete the experiment due to numerical instability, further underscoring the stability and effectiveness of our approach in handling real-world label noise. These consolidated results validate the strong performance of LENL-NS for learning with noisy labels.

Performance on CIFAR-100N Dataset. On the CIFAR-100N benchmark, LENL-NS achieves 61.51% accuracy, outperforming all baselines, including PLM (60.52%) and CE (55.50%). The 6% improvement over the vanilla cross-entropy baseline (CE) demonstrates that explicitly regularizing the transition matrix's effective rank significantly enhances robustness to real-world label noise.

Performance on Food-101N Dataset. On the Food-101N benchmark, LENL-NS achieves state-of-the-art accuracy of 83.30%. This significantly outperforms the standard Cross-Entropy (CE) baseline (80.34%) and the recent PLM (80.89%).

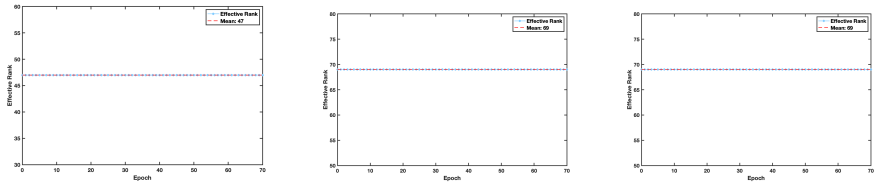
Table 5: Comparison of accuracy, precision, recall, and F1 score on the Noisy Ostracods dataset and accuracy on the CIFAR-100N and Food-101N datasets. The best performance is indicated in bold. '-' means that the method fails to complete due to numerical instability. Standard deviations are not reported on Noisy Ostracods following the original papers [9].

Methods/Datasets	Noisy Ostracods				CIFAR-100N	Food-101N
	Accuracy	Precision	Recall	F1	Accuracy	Accuracy
CE [5]	95.98	**88.50**	77.80	79.51	55.50 ± 0.66	80.34
VolMinNet [17]	-	-	-	-	57.80 ± 0.31	-
Co-Teaching [7]	95.79	79.01	71.79	73.23	57.88 ± 0.24	80.59
PLM [45]	96.77	77.77	72.74	73.46	60.52 ± 0.32	80.89
ILDE [18]	96.23	77.49	72.08	73.05	57.83 ± 0.61	76.06
LENL-NS (Ours)	**97.20**	87.21	**77.93**	**79.75**	**61.51 ± 0.14**	**83.30**

4.3 Reduced Effective Rank Verification

Figure 5 shows the effective rank of the transition matrix $\hat{T}$ varying with epoch number.

The effective rank of the estimated transition matrix $\hat{T}$ remains consistently low relative to the exact rank throughout the training process. This observation confirms that the reduced effective-rank property persists throughout training, indicating that the reduced effective-rank property is actively preserved by nuclear-norm regularization.



(a) Noisy Ostracods **(b)** CIFAR-100 (Sym-20 %) **(c)** CIFAR-100 (Sym-50%)

Fig. 5: Effective Rank vs. Epoch on various datasets for our estimated $\hat{T}$.

4.4 Extra Experiments

We conduct additional experiments analyzing the evolution of T's rank during training, which reveal that T remains non-degenerate during training. We also conduct ablation studies. Due to space limitations, these results are presented in Appendix **??**.

5 Conclusion

We present the first systematic study revealing that noise transition matrices exhibit reduced effective rank. Leveraging this, we propose nuclear norm regu-

larization to encourage this structure and employ Newton–Schulz iteration to compute the regularizer stably, avoiding the numerical instability of SVD-based methods. Across CIFAR-100, CIFAR-100N, ImageNet-1K, and Noisy Ostracods, LENL delivers substantial accuracy gains over all competitors, confirming that low-effective-rank modeling is powerful in practice.

6 Future Work

While LENL effectively exploits low effective rank structure in noise transition matrices, it assumes all samples are in-distribution. Real-world datasets contain out-of-distribution samples mislabeled as in-distribution classes. These outliers lack semantic proximity to any class cluster, yielding full-rank mislabeling patterns that violate the low-rank assumption. Integrating outlier detection while preserving low-rank constraints thus remains a promising direction.

Acknowledgment

This work was supported by the Scientific Research Start-up Fund of Jilin University (Grant No.419080526A65), the Natural Science Foundation of Fujian Province (Grant 2026J008138) and the Science Foundation for Youth of the Education Department of Fujian Province, China (Grant JZ250007). We would like to thank Lei Feng for guiding Runsheng Rademacher's complexity. We also thank Suming Yu for her assistance.

References

1. Algan, G., Ulusoy, I.: Meta soft label generation for noisy labels. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 7142–7148. IEEE (2021)
2. Cheng, D., Liu, T., Ning, Y., Wang, N., Han, B., Niu, G., Gao, X., Sugiyama, M.: Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16630–16639 (2022)
3. Cheng, J., Liu, T., Ramamohanarao, K., Tao, D.: Learning with bounded instance and label-dependent label noise. In: International conference on machine learning. pp. 1789–1799. PMLR (2020)
4. Dai, D., Zhu, H., Xia, S., Wang, G.: Granular-ball representation learning for deep cnn on learning with label noise. In: International Conference on Neural Information Processing. pp. 44–58. Springer (2024)
5. De Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Annals of operations research **134**(1), 19–67 (2005)
6. Goldberger, J., Ben-Reuven, E.: Training deep neural-networks using a noise adaptation layer. In: International conference on learning representations (2017)
7. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems **31** (2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision. pp. 630–645. Springer (2016)
9. Hu, J., Yuanyuan, H., Chen, Y., Wang, H., Yasuhara, M.: Noisy ostracods: A fine-grained, imbalanced real-world dataset for benchmarking robust machine learning and label correction methods. Advances in Neural Information Processing Systems **37**, 50750–50771 (2024)
10. Huang, B., Lin, Y., Xu, C.: Contrastive label correction for noisy label learning. Information sciences **611**, 173–184 (2022)
11. Ionescu, C., Vantzos, O., Sminchisescu, C.: Matrix backpropagation for deep networks with structured layers. In: Proceedings of the IEEE international conference on computer vision. pp. 2965–2973 (2015)
12. Jolliffe, I.T.: Principal component analysis. Springer, 2nd edn. (2002)
13. Karim, N., Rizve, M.N., Rahnavard, N., Mian, A., Shah, M.: Unicon: Combating label noise through uniform selection and contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9676–9686 (2022)
14. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
15. Lee, K.H., He, X., Zhang, L., Yang, L.: Cleannet: Transfer learning for scalable image classifier training with label noise. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5447–5456 (2018)
16. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394 (2020)
17. Li, X., Liu, T., Han, B., Niu, G., Sugiyama, M.: Provably end-to-end label-noise learning without anchor points. In: International conference on machine learning. pp. 6403–6413. PMLR (2021)
18. Liao, Z., et al.: Instance-dependent label distribution estimation for learning with label noise. IJCV (2025)

19. Lienen, J., Hüllermeier, E.: From label smoothing to label relaxation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 8583–8591 (2021)
20. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence **38**(3), 447–461 (2015)
21. Liu, Y., Guo, H.: Peer loss functions: Learning from noisy labels without knowing noise rates. In: International conference on machine learning. pp. 6226–6236. PMLR (2020)
22. Ortego, D., Arazo, E., Albert, P., O'Connor, N.E., McGuinness, K.: Multi-objective interpolation training for robustness to label noise. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6606–6615 (2021)
23. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1944–1952 (2017)
24. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM review **52**(3), 471–501 (2010)
25. Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596 (2014)
26. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
27. Schulz, G.: Iterative berechung der reziproken matrix. ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik **13**(1), 57–59 (1933)
28. Singhal, A., et al.: Modern information retrieval: A brief overview. IEEE Data Eng. Bull. **24**(4), 35–43 (2001)
29. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. IEEE transactions on neural networks and learning systems **34**(11), 8135–8153 (2022)
30. Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5552–5560 (2018)
31. Udell, M., Townsend, A.: Why are big data matrices approximately low rank? SIAM Journal on Mathematics of Data Science **1**(1), 144–160 (2019)
32. Van Rooyen, B., Menon, A., Williamson, R.C.: Learning with symmetric label noise: The importance of being unhinged. Advances in neural information processing systems **28** (2015)
33. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 322–330 (2019)
34. Wei, H., Feng, L., Chen, X., An, B.: Combating noisy labels by agreement: A joint training method with co-regularization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13726–13735 (2020)
35. Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., Liu, Y.: Learning with noisy labels revisited: A study using real-world human annotations. arXiv preprint arXiv:2110.12088 (2021)

36. Wu, S., Xia, X., Liu, T., Han, B., Gong, M., Wang, N., Liu, H., Niu, G.: Class2simi: A noise reduction perspective on learning with noisy labels. In: International conference on machine learning. pp. 11285–11295. PMLR (2021)
37. Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., Chang, Y.: Robust early-learning: Hindering the memorization of noisy labels. In: International conference on learning representations (2020)
38. Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., Sugiyama, M.: Part-dependent label noise: Towards instance-dependent label noise. Advances in neural information processing systems **33**, 7597–7610 (2020)
39. Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., Sugiyama, M.: Are anchor points really indispensable in label-noise learning? Advances in neural information processing systems **32** (2019)
40. Yao, Y., Sun, Z., Zhang, C., Shen, F., Wu, Q., Zhang, J., Tang, Z.: Jo-src: A contrastive approach for combating noisy labels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5192–5201 (2021)
41. Yao, Y., Liu, T., Han, B., Gong, M., Deng, J., Niu, G., Sugiyama, M.: Dual t: Reducing estimation error for transition matrix in label-noise learning. Advances in neural information processing systems **33**, 7260–7271 (2020)
42. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., Sugiyama, M.: How does disagreement help generalization against label corruption? In: International conference on machine learning. pp. 7164–7173. PMLR (2019)
43. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530 (2016)
44. Zhang, W., Wang, Y., Qiao, Y.: Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7373–7382 (2019)
45. Zhao, R., Shi, B., Ruan, J., Pan, T., Dong, B.: Estimating noisy class posterior with part-level labels for noisy label learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22809–22819 (2024)