COVERAGE RISK MINIMIZATION: A UNIFIED FRAMEWORK FOR LONG-TAILED DEEP FINE-GRAINED CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-tailed distributions and fine-grained semantics frequently co-occur in real-world visual data: tail classes lack sufficient data support, while semantically similar categories share highly similar visual representations, rendering reliable intra-class aggregation and inter-class separation considerably more challenging. To this end, we revisit fine-grained long-tailed clustering from the perspective of forward representation optimization and propose **D4C**, a hierarchical coverage risk optimization framework with density-aware forward control that continuously shapes clustering-conducive geometric structures across different representation levels. Theoretically, starting from the PAC upper bound of coverage risk, we isolate an optimizable average term and a geometric bottleneck governed by the worst-case covering radius, and adopt optimal transport and geometric repulsive potentials as their differentiable surrogates to coordinate sample coverage and center separation. Concurrently, we introduce a vMF mixture model at each representation level to characterize the local density structure, guiding samples to aggregate toward local density peaks. To tackle imbalanced support among long-tailed categories, we further design long-tail-aware transport marginals, and over-segmentation with Ward merging and EMA teacher freezing keep coverage-risk minimization stable. Extensive experiments on multiple long-tailed and fine-grained clustering benchmarks demonstrate that D4C achieves state-of-the-art performance.

1 INTRODUCTION

Deep clustering aims to learn discriminative representations from unlabeled visual data and discover underlying category structures(Zhou et al., 2024). With the advancement of visual pre-trained models(Grill et al., 2020; Caron et al., 2020), the representational capacity of deep clustering has been substantially enhanced, yet long-tailed distributions(Li & Jia, 2025; Jiang et al., 2021) and fine-grained semantics(Wei et al., 2021) in real-world data still pose formidable challenges. On the one hand, tail classes are supported by only a limited number of samples, making their statistical structures difficult to estimate reliably. On the other hand, fine-grained categories typically exhibit highly similar visual representations and can be distinguished solely by subtle local differences. When these two characteristics coexist, imbalanced category support and limited inter-class variations compound each other, making different categories more prone to entanglement in the feature space, thereby substantially increasing the difficulty of unsupervised category discovery(Hoang et al., 2025; Yi et al., 2025).

Existing deep clustering methods typically leverage mechanisms such as pseudo-labeling(Caron et al., 2018), self-training(Van Gansbeke et al., 2020), contrastive learning(Li et al., 2021), or prototype learning to optimize the clustering relationships among samples in the final representation space. However, for fine-grained long-tailed data, merely characterizing the clustering structure in the final representation is insufficient to explicitly capture geometric properties such as sample coverage, local density, and cluster center separation(Villani et al., 2009) throughout the forward representation process. This raises a more fundamental question: **How to continuously construct a clustering geometry conducive to category discovery as representations evolve?** We argue that a reliable intermediate representation should simultaneously satisfy two fundamental conditions:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

samples must be sufficiently covered by their corresponding cluster centers, while distinct centers must maintain necessary geometric separation. Motivated by this insight, we approach deep clustering from the perspective of forward representation optimization and reformulate it as a **hierarchical coverage risk optimization** problem.

To this end, we propose **Density-Driven Determinantal Deep Clustering (D4C)**, a hierarchical coverage risk optimization framework. D4C extends the conventional single-head design into layer-wise control over the forward representations, continuously optimizing sample coverage and clustering geometry across different representation levels. Theoretically, starting from the PAC upper bound of coverage risk, we decompose it into an average term reflecting overall coverage quality and a worst-case term characterizing extreme coverage scenarios, and construct corresponding differentiable optimization objectives for each. For the mean term, we establish a global coupling between samples and cluster centers via optimal transport(Cuturi, 2013) to coordinate sample coverage and center support. Furthermore, addressing the mismatch between uniform marginals and true category support under long-tailed distributions, we design long-tail-aware transport marginals, enabling distinct centers to achieve transport quality commensurate with the data distribution. For the worst-case term, we further introduce a geometric repulsive potential(Kulesza & Taskar, 2012) to explicitly constrain inter-center distances, suppressing center redundancy and over-aggregation, thereby enhancing the geometric separability between adjacent categories. Together, they constitute the global geometric constraints for hierarchical coverage risk optimization.

Building upon this foundation, D4C simultaneously models the local density structures across different representation levels. Specifically, we introduce a vMF mixture model(Banerjee et al., 2005) at each hidden layer to estimate local density peaks based on the current feature distribution. Guided by posterior membership, samples are directed to aggregate toward their corresponding high-density regions, thereby promoting intra-class compactness and preserving the local peak-valley structure. Consequently, optimal transport and geometric repulsion are responsible for shaping global coverage and separation, while local density modeling further constrains how samples aggregate around latent category structures. Together, they constitute density-aware hierarchical coverage optimization. Furthermore, to accommodate complex, imbalanced category structures, over-segmentation with Ward merging(Hartigan & Wong, 1979) provides ample prototype capacity so that coverage risk can be minimized without conflating adjacent structures, while EMA teacher freezing(Siméoni et al., 2025) maintains a stable reference for this minimization.

Our contributions are as follows:

1. We propose a novel hierarchical coverage risk optimization framework, D4C, which continuously shapes clustering-conducive geometric structures across multiple representation levels.
2. Starting from the PAC upper bound of coverage risk, we establish its theoretical connections to optimal transport and geometric repulsive potentials. Furthermore, we propose long-tail-aware transport marginals to accommodate data distributions characterized by highly imbalanced category support.
3. We integrate multi-layer vMF density modeling, over-segmentation and Ward merging, and an EMA teacher freezing mechanism to further enhance local structure modeling and training stability. On multiple long-tailed and fine-grained clustering benchmarks, D4C achieves state-of-the-art performance, and systematic analyses validate the effectiveness of each constituent module.

2 Related Work

Deep Clustering under Long-Tailed Distributions. Deep clustering jointly learns representations and cluster assignments without annotations Xie et al. (2016); Caron et al. (2018); Zhou et al. (2024). Recent advances primarily rely on contrastive learning and prototype-based methods to construct clustering-friendly features Van Gansbeke et al. (2020); Li et al. (2021); Tao et al. (2021); Li & Jia (2025). When confronted with long-tailed distributions, head classes tend to dominate the feature space, causing tail structures to degrade Jiang et al. (2021); Li et al. (2025a). Existing unsupervised long-tailed clustering methods typically address this by employing contrastive representations with reweighting strategies Zhou et al. (2022); Hoang et al. (2025) or introducing auxiliary clusters to rebalance the pseudo-labels Yi et al. (2025); Li et al. (2025a). However, these methods predominantly perform rebalancing and structural correction exclusively in the final embedding space.

For fine-grained long-tailed data, where inter-class variations are subtle, relying solely on the final representation is fundamentally insufficient. The discriminative cues for tail classes are often overwhelmed and collapsed during the early and middle stages of forward propagation. Rebalancing at the final layer cannot recover the fine-grained geometric structures that have already been lost. D4C departs from this paradigm by correcting the layer-wise coverage geometry continuously throughout the forward representation process, ensuring that fine-grained tail structures are preserved from shallow to deep layers.

Unsupervised Fine-Grained Clustering. Fine-grained clustering aims to discover visually similar subordinate categories with low inter-class and high intra-class variance Wei et al. (2021). While early explorations relied on generative or hierarchical models Ji et al. (2019); Singh et al. (2019); Kim & Ha (2021), recent progress has been driven by contrastive and prototype-based unsupervised methods Metaxas et al. (2023); Liu et al. (2024); Adaloglou et al. (2023). Some concurrent works also attempt to leverage external knowledge from Vision-Language Models (VLMs) to guide the clustering process Radford et al. (2021); Peng et al. (2026); Ye et al. (2026). Despite these advances, unsupervised fine-grained clustering remains highly challenging, especially under long-tailed settings. Existing methods typically lack explicit mechanisms to control the hierarchical geometric structures (e.g., local density peaks and inter-center separation) across different representation levels. Consequently, they are prone to mode collapse, where subtle tail categories are merged into dominant head clusters. Unlike methods that rely on external VLM priors or merely optimize the final feature space, D4C targets pure unsupervised fine-grained clustering and shapes layer-wise coverage geometry to explicitly preserve local density and global separation.

3 METHODS

Problem Formulation. Given unlabeled data $\{x_i\}_{i=1}^N$, the goal is to partition them into the underlying categories, recovered at evaluation as K^* groups by Ward merging (Sec. 3.4). Starting from normalized features $f_i^{(0)}$ produced by a fixed pre-trained encoder(Grill et al., 2020; Caron et al., 2020; Siméoni et al., 2025), D4C trains an L-layer forward stack h_θ so that each layer is covered by cluster centers and the final layer yields cluster assignments, whose l-th layer outputs:

$$f_i^{(l)} = \frac{h_l(f_i^{(l-1)})}{\|h_l(f_i^{(l-1)})\|_2}, \quad l = 1, \dots, L-1. \tag{1}$$

Layer-wise, D4C minimizes the coverage risk through four complementary mechanisms: (1) OT average coverage (Sec. 3.1) coordinates sample coverage and center support; (2) geometric repulsion (Sec. 3.1) prevents center degeneracy and bounds the worst-case radius; (3) vMF density pulling (Sec. 3.2) draws features toward local density peaks for intra-class compactness; and (4) over-segmentation with EMA freezing (Sec. 3.4) provides sufficient prototype capacity and a stable optimization reference. These are detailed below. An overview of the complete pipeline is illustrated in Fig. 1.

3.1 COVERAGE CONTROL VIA OPTIMAL TRANSPORT AND GEOMETRIC REPULSION

The layer-wise coverage risk is $R_\tau(C) = \Pr[\min_m \|x - c_m\| > \tau]$, where x denotes the current layer's features $f^{(l)}$ (layer index omitted), i.e., the probability that a sample falls outside the τ-neighborhood of any center. Minimizing it ensures that the layer's representation can be linearly separated by the nearest center within radius τ: samples lie within a center's neighborhood and each center has support. When both marginals hold, this optimization is realized by optimal transport(Cuturi, 2013; Villani et al., 2009), which couples samples to centers under marginal constraints.

Represent the empirical samples and the peak set as discrete distributions $\hat{\mu}_N = \frac{1}{N}\sum_i \delta_{x_i}$ and $\nu_C = \sum_m \omega_m \delta_{c_m}$, where ω_m are the vMF mass weights. For any coupling Π satisfying the marginals and for the l-th layer (omitting the layer index),

$$\sum_{i,m} \Pi_{im} \|x_i - c_m\|^2 \;\geq\; \frac{1}{N} \sum_i \min_m \|x_i - c_m\|^2 = \Psi(C). \tag{2}$$

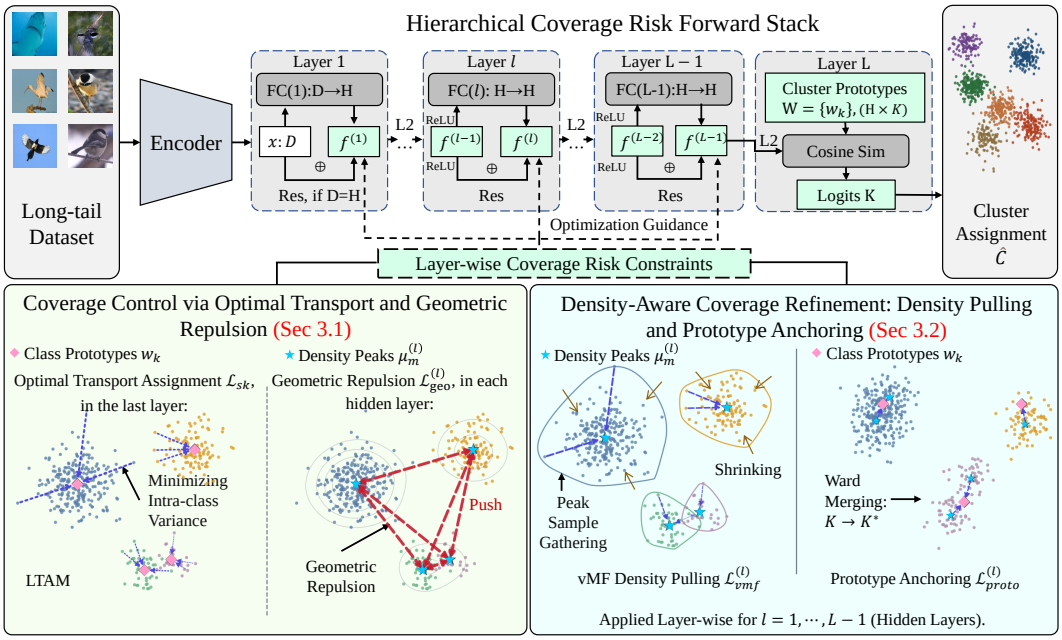


Figure 1: **Overview of D4C.** A multi-layer forward stack internalizes clustering objectives layer by layer: optimal transport covers the PAC mean term $\mathcal{L}_{\text{sk}}$(in last layer), geometric repulsion bounds the worst-case term $\mathcal{L}_{\text{geo}}^{(l)}$(in hidden layers), vMF density pulling $\mathcal{L}_{\text{vmf}}^{(l)}$ pulls samples toward local density peaks, and prototype anchoring $\mathcal{L}_{\text{proto}}^{(l)}$ aligns hidden peaks to class prototypes. Over-segmented prototypes are merged by Ward clustering at evaluation. LTAM denotes long-tail-aware marginals.

Here $\Psi(C)$ is the average squared distance from each sample to its nearest center (the unconstrained baseline cost). Taking the infimum over Π on both sides yields $\Psi(C) \leq W_2^2(\hat{\mu}_N, \nu_C)$. Thus minimizing the transport cost minimizes an upper bound on $\Psi(C)$; we adopt it as a differentiable proxy for the average coverage risk. The bound is tight whenever the hard nearest-center coupling satisfies the marginals, in which case $\Psi(C) = W_2^2(\hat{\mu}_N, \nu_C)$ holds. In hidden layers this proxy is the per-batch posterior of Sec. 3.2; in the output layer it is the Sinkhorn assignment of Appendix F. For the last layer, the OT cost $-\langle f, w_k \rangle$ is affinely equivalent to the squared distance $\|f - w_k\|^2$ on the unit sphere, since $\|f - w_k\|^2 = 2 - 2\langle f, w_k \rangle$; replacing centers with class prototypes then gives the risk-minimization loss $\mathcal{L}_{\text{sk}}$, whose Sinkhorn target t_{ik} is computed below. For $\mathcal{L}_{\text{sk}}$ we use two augmented views per sample: the Sinkhorn target t_{ik} is computed on the stop-gradient view's last hidden representation $\tilde{f}_i^{(L-1)}$ against the EMA teacher prototypes $\bar{W}$ (Sec. 3.4), and supervises the gradient view's class predictions. The explicit definition of $\mathcal{L}_{\text{sk}}$ is in Appendix F.

Long-tail-aware marginals are applied only to the final-layer risk objective, since the long-tail prior is class-level; for hidden layers, the OT marginals are peak mass distributions and require no correction. Under long tails, the uniform marginal $m_k = 1/K$ is mathematically feasible but statistically mismatched(Zhang et al., 2024; Shi et al., 2024) with the true class distribution: each class is forced to absorb a $1/K$ fraction of total mass, whereas tail classes have far smaller sample proportions, forcing tail prototypes to absorb head samples and deviate from their true data centroids, thereby inflating the coverage radius. We relax the marginal to $m_k \propto \pi_k^{\alpha}, \quad \alpha \in [0, 1]$, where π_k is the empirical class prior. Under over-segmentation, $\pi_k = (1-\eta)\,\hat{\pi}_k + \eta/K$ with $\hat{\pi}_k$ the fuzzy C-means (FCM) pseudo-class mass on the $K = \rho K^*$ prototypes from warm-up and $\eta = 0.5$; in the no-over-segmentation ablation ($K = K^*$), $\pi_k \propto r^{-k/(K-1)}$, a geometric prior implied by the imbalance ratio r (Sec. 4.1). Setting $\alpha = 0$ recovers the uniform marginal, and $\alpha = 1$ fully follows the prior. This interpolation alleviates statistical mismatch while retaining some balancing ability.

Furthermore, if several centers nearly coincide, the worst-case coverage radius $r_{\max}(C)$ (Eq. 11) remains high, violating the risk control objective. This necessitates introducing a geometric repulsive potential on top of the transport cost. Define the similarity matrix $S^{(l)}$, where the centers are the

layer's density peaks, $c_m = \mu_m^{(l)}$,

$$S_{ij}^{(l)} = \exp\Big(-\frac{\|\mu_i^{(l)} - \mu_j^{(l)}\|^2}{2\sigma^2}\Big), \tag{3}$$

where maximizing $\det(S^{(l)})$ effectively enlarges inter-center distances (strictly monotonic in the equiangular case, see Appendix D). Taking its negative logarithm as a regularizer:

$$\mathcal{L}_{\text{geo}}^{(l)} = -\frac{1}{M^{(l)}} \log\det(S^{(l)} + \gamma I), \tag{4}$$

where $\gamma > 0$ is a small numerical jitter ensuring positive definiteness. This regularizer maximizes the effective volume of centers on the hypersphere, preventing degeneracy in the OT solution.

3.2 Density-Aware Coverage Refinement: Density Pulling and Prototype Anchoring

Density peak discovery. Coverage risk minimization also requires samples to stay close to their centers. We employ a vMF mixture model(Banerjee et al., 2005) at each hidden layer to estimate the local density structure and softly pull samples toward their assigned peaks. For the l-th hidden layer features $F^{(l)} = [f_1^{(l)}, \ldots, f_N^{(l)}]$, we model the density structure using a vMF mixture:

$$p(f) = \sum_{m=1}^{M^{(l)}} \omega_m^{(l)} C_d(\kappa_m^{(l)}) \exp\big(\kappa_m^{(l)} \langle f, \mu_m^{(l)} \rangle\big), \tag{5}$$

where $\mu_m^{(l)}$ is the peak mean direction, $\kappa_m^{(l)}$ is the concentration parameter, and $\omega_m^{(l)}$ is the mixture weight. Peak parameters are re-estimated via EM after feature updates; the number of peaks $M^{(l)}$ follows the over-segmentation rule in Appendix F, and low-mass peaks are filtered, so head classes may form multiple peaks while sparse tail classes share peaks. This yields the posterior probability:

$$q_{im}^{(l)} = \frac{\omega_m^{(l)} C_d(\kappa_m^{(l)}) \exp\big(\kappa_m^{(l)} \langle f_i^{(l)}, \mu_m^{(l)} \rangle\big)}{\sum_{m'=1}^{M^{(l)}} \omega_{m'}^{(l)} C_d(\kappa_{m'}^{(l)}) \exp\big(\kappa_{m'}^{(l)} \langle f_i^{(l)}, \mu_{m'}^{(l)} \rangle\big)}. \tag{6}$$

Given peak parameters $(\mu_m, \kappa_m, \omega_m)$, the normalized posterior $q_{im}^{(l)}/N$ is the ε=1 entropy-regularized optimal transport plan between the sample distribution and the peak-mass distribution $\nu_C = \sum_m \omega_m \delta_{\mu_m}$ under the cost $-\kappa_m \langle f_i^{(l)}, \mu_m^{(l)} \rangle$, which is affinely equivalent to the squared distance on the sphere (Proposition in Appendix C). It therefore instantiates the same cost as the optimal transport objective of Sec. 3.1, acting on local density structure rather than on global support.

Feature pulling toward peaks. To suppress the monopoly of dominant peaks over gradient directions, we introduce a square-root weighting of peak masses:

$$\tilde{\pi}_m^{(l)} \propto \Big(\sum_{i=1}^{N} q_{im}^{(l)}\Big)^{1/2}, \quad \sum_{m=1}^{M^{(l)}} \tilde{\pi}_m^{(l)} = 1. \tag{7}$$

The density pulling loss softly pulls hidden features toward the mean directions of their assigned peaks using posterior weighting:

$$\mathcal{L}_{\text{vmf}}^{(l)} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{m=1}^{M^{(l)}} \tilde{\pi}_m^{(l)} q_{im}^{(l)} \kappa_m^{(l)} \langle f_i^{(l)}, \mu_m^{(l)} \rangle. \tag{8}$$

The gradient $\nabla_{f_i} \mathcal{L}_{\text{vmf}}^{(l)} = -\frac{1}{N} \sum_m \tilde{\pi}_m q_{im} \kappa_m \mu_m$. drives features to migrate softly toward local high-density regions. To stabilize training, we keep $\kappa_m = 1$ in Eq. 8.

Prototype anchoring. To propagate the density structure to the classification layer, we align the final-layer class prototypes w_k with the hidden peaks:

$$\mathcal{L}_{\text{proto}}^{(l)} = -\sum_{m=1}^{M^{(l)}} \tilde{\pi}_m^{(l)} \sum_{k=1}^{K} a_{mk}^{(l)} \big\langle \mu_m^{(l)}, w_k \big\rangle, \quad a_{mk}^{(l)} = \text{softmax}_k\Big(\frac{\langle \mu_m^{(l)}, w_k \rangle}{\tau_p}\Big). \tag{9}$$

3.3 PAC Upper Bound of Coverage Risk

Having detailed the algorithmic components, we now formalize their theoretical foundation via a PAC upper bound (Vapnik, 1999): OT minimizes an upper bound on $\Psi(C)$, while the geometric potential keeps $r_{\max}(C)$ non-degenerate. Let the features of a given layer be random points $x \sim P$ on the unit sphere $\mathbb{S}^{d-1}$, and the candidate center set be $C = \{c_1, \ldots, c_M\} \subset S^{d-1}$, instantiated as $\{\mu_m^{(l)}\}$ for a hidden layer or $\{w_k\}$ for the output layer; we omit the layer index l for brevity. For the coverage risk $R_\tau(C)$ defined earlier (the probability that a sample falls outside the τ-neighborhood of any center), its $0/1$ form is non-differentiable. Using Markov's and Hoeffding's inequalities(Hoeffding, 1963), we obtain, with probability at least $1-\delta$ over the training sample, the following optimizable surrogate (full derivation in Appendix A):

$$R_\tau(C) \le \frac{1}{\tau^2}\Big(\underbrace{\frac{1}{N}\sum_i \min_m \|x_i - c_m\|^2}_{\Psi(C)} + 4\sqrt{\frac{\ln(1/\delta)}{2N}} \Big). \tag{10}$$

Only the nearest-peak cost $\Psi(C)$ is optimizable; the $O(N^{-1/2})$ sample-complexity term is fixed, so Secs. 3.1–3.2 target $\Psi(C)$ only. Since our centers are estimated from the same data, Appendix B upgrades this fixed-center bound to an adaptive one via uniform convergence over the center set, adding an $O(\sqrt{dM/N})$ complexity penalty while keeping the same optimizable leading term $\Psi(\hat{C})$.

$$r_{\max}(C) = \max_i \min_m \|x_i - c_m\|, \qquad \Psi(C) \le r_{\max}^2(C). \tag{11}$$

Since $\Psi(C) \le r_{\max}^2(C)$ (Eq. 11), reducing $r_{\max}(C)$ tightens the dominant term of Eq. 10. Accordingly, OT controls $\Psi(C)$, while the geometric potential keeps $r_{\max}(C)$ non-degenerate; this separation transfers to $\{w_k\}$ through anchoring (Lemma E, Appendix E). The vMF pull does not change the form of the bound; it tightens the bound's inputs by reducing the empirical nearest-peak cost $\Psi(C)$ and the radius $r_{\max}(C)$, so the three objectives act jointly on global coupling, global separation, and local compactness (ablated in Table 5).

3.4 EMA and Over-segmentation

Over-segmentation. Under the true number of classes K^*, adjacent yet separable class pairs are often forced to share a single prototype, causing insufficient prototype capacity. Therefore, D4C trains with $K = \rho K^*$ over-segmented prototypes, where ρ is a hyperparameter ($\rho = 2$) that provides excess prototype capacity so that adjacent yet separable class pairs need not share a single prototype. At evaluation, the K prototypes $W \in \mathbb{R}^{H\times K}$ (columns w_k) are merged into K^* groups via Ward hierarchical clustering (Ward Jr, 1963); K^* is the ground-truth class count, used only for metric computation:

$$\Delta(A,B) = \frac{|A|\cdot|B|}{|A|+|B|}\|\widehat{w}_A - \widehat{w}_B\|_2^2, \qquad \widehat{w}_A = \tfrac{1}{|A|}\sum_{k\in A} w_k. \tag{12}$$

Geometric repulsion (Eq. 4) encourages dispersed density peaks, and prototype anchoring (Eq. 9) aligns peaks with class prototypes, providing the structure that Ward merging consolidates.

EMA teacher freezing. Target drift in self-training can destabilize training and degrade the learned prototypes(Asano et al., 2019; Caron et al., 2020). To prevent this, we maintain an EMA teacher prototype $\bar{W}_t = \rho_e \bar{W}_{t-1} + (1-\rho_e)W_t$, updated only in the early training stage; after freezing we drop the subscript and write $\bar{W}$. This provides a stable reference frame for OT risk minimization and density peaks.

3.5 Joint Optimization

The risk terms above are jointly minimized via the following objective: the first term is the risk mean term of the final layer, a cross-entropy under long-tail-aware marginals; the remaining terms are layer-wise risk constraints.

$$\mathcal{L} = \mathcal{L}_{\text{sk}} + \lambda_{\text{hid}} \sum_{l=1}^{L-1} \big(\mathcal{L}_{\text{geo}}^{(l)} + \mathcal{L}_{\text{vmf}}^{(l)}\big) + \beta_{\text{hid}} \sum_{l=1}^{L-1} \mathcal{L}_{\text{proto}}^{(l)}. \tag{13}$$

Table 1: Comparison of clustering performance on three fine-grained datasets with imb_ratio = 10. Best results are in bold; second-best results are underlined.

Methods	CUB-200-2011				Stanford Dogs				Oxford Flowers			
	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI
IDC (NeurIPS2024)	14.8	14.7	26.2	0.8	5.5	8.7	17.0	0.0	13.6	13.6	22.6	1.9
CLUDI (ICML2025)	19.8	19.7	59.2	5.6	22.8	22.6	53.7	11.8	45.3	44.9	67.5	29.6
LFSS (ICML2025)	24.6	24.7	56.9	11.9	29.5	26.6	53.0	19.3	50.4	41.2	67.3	39.8
ConMix (ICLR2025)	22.3	22.2	55.8	10.1	30.6	36.5	51.8	19.8	52.2	49.1	68.4	**49.3**
MiniClustering (ICLR2026)	26.5	25.9	59.2	13.3	36.8	33.6	59.7	21.5	37.8	36.7	59.1	13.4
Ours	**28.6**	**26.3**	**61.2**	**17.1**	**44.0**	**37.3**	**64.7**	**27.0**	**53.0**	**51.2**	**73.2**	47.2

Table 2: Performance comparison on larger datasets with an imb_ratio = 10. Best results are in bold; second-best results are underlined.

Methods	Tiny-ImageNet-LT				ImageNet-100-LT				ImageNet-50-LT			
	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI
LFSS (ICML2025)	22.9	18.4	28.7	14.0	54.3	52.4	76.2	43.1	29.6	26.2	52.3	18.4
ConMix (ICLR2025)	16.5	14.5	34.9	8.3	37.3	35.3	64.7	26.6	29.1	25.9	50.4	16.2
MiniClustering (ICLR2026)	25.5	19.4	40.6	14.9	52.6	45.8	74.4	41.9	33.9	26.7	53.7	22.6
Ours	**49.9**	**46.8**	**70.4**	**40.3**	**60.6**	**58.6**	**80.2**	**50.0**	**38.8**	**35.5**	**57.5**	**30.6**

where $\mathcal{L}_{\text{sk}}$ is the long-tail-marginal cross-entropy defined in Appendix F; the three regularizers correspond to Eqs. 4, 8, and 9, with weights λ_{hid} and β_{hid} controlling their strengths. The complete procedure is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Settings

Datasets. Following standard long-tailed deep clustering protocols(Li & Jia, 2025; Li et al., 2025a), we evaluate D4C on six long-tailed benchmarks—CIFAR-10, CIFAR-20(Krizhevsky et al., 2009) and STL-10(Coates et al., 2011) with imbalance ratios 5 and 10, plus Tiny-ImageNet-LT(Le et al., 2015), ImageNet-100-LT and ImageNet-50-LT(Deng et al., 2009; Liu et al., 2019) with imbalance ratio 10—and on three fine-grained long-tailed benchmarks, CUB-200-2011(Wah et al., 2011), Stanford Dogs(Khosla et al., 2011) and Oxford Flowers(Nilsback & Zisserman, 2008). CIFAR-20 uses the 20 super-classes of CIFAR-100 as labels, and ImageNet-100/50 are randomly sampled ImageNet subsets. Long-tailed training sets are constructed by exponential sub-sampling with imbalance ratio $r = \max_k N_k / \min_k N_k$(Cao et al., 2019; Zhou et al., 2020), and test sets are truncated to the same distribution; following the transductive protocol of ConMix(Li & Jia, 2025), all models are trained and evaluated on the same images and class distribution. On STL-10 we combine the official train and test splits for unsupervised training(Tao et al., 2021), and the fine-grained datasets follow the same long-tailed construction with imb_ratio = 10. Note that CAA denotes the mean per-class recall after Hungarian matching (Kuhn, 1955).

Implementation details. We use a ResNet-18 encoder(He et al., 2016) with the publicly released SimCLR weights(Chen et al., 2020) from the SCAN repository(Van Gansbeke et al., 2020), and keep it frozen. D4C trains a two-layer forward stack with 512-d hidden units ($L = 2$, i.e., one hidden layer; deeper stacks are studied in Fig. 2d). A vMF mixture is re-estimated by EM in each hidden layer, and the final-layer coverage loss uses long-tail-aware marginals. Training uses over-segmented prototypes ($\rho = 2$) merged by Ward clustering (Ward Jr, 1963) at evaluation, with the EMA teacher frozen after the early stage ($\xi = 0.6$). The joint objective (Eq. 13) is optimized with Adam(Kingma & Ba, 2014) (batch size 512) for 200 epochs, with $\lambda_{\text{hid}} = 0.05$, $\beta_{\text{hid}} = 0.2$ and $\alpha = 0.4$; wider hidden layers are used for fine-grained datasets, and the remaining fine-grained settings are listed in the Appendix F. For large-scale benchmarks we use DINOv3 weights(Siméoni et al., 2025) with all other hyperparameters at the default settings above, and for fine-grained datasets we use MoCo v3(Chen et al., 2021) ResNet-50. The remaining implementation details are given in Appendix F.

Table 3: Clustering performance in percentage (%) on three benchmark datasets (imb_ratio = 5) across various methods. Best results are in bold; second-best results are underlined.

Method	CIFAR-10				CIFAR-20				STL-10			
	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI
SCAN (ECCV2020)	55.6	60.3	54.0	39.3	38.8	39.7	42.6	25.1	48.8	49.8	46.2	34.9
SDCLR (ICML2021)	44.1	50.5	43.4	38.0	40.2	38.7	40.3	23.8	35.8	37.4	34.4	19.3
CC (AAAI2021)	25.2	20.9	18.3	1.47	16.0	12.0	18.0	0.00	41.5	30.0	46.2	22.4
IDFD (ICLR2021)	56.7	63.0	51.8	37.1	31.2	30.7	30.5	16.2	41.6	41.2	39.6	24.3
CoNR (NeurIPS2023)	41.4	43.5	34.9	23.3	23.8	23.4	21.6	12.1	32.4	31.3	29.0	18.2
ProPos (TPAMI2022)	51.4	59.2	52.2	34.1	40.7	39.3	42.6	26.4	37.2	39.3	34.5	21.8
DMICC (AAAI2023)	40.6	42.5	36.9	25.4	25.2	23.1	20.6	7.5	45.9	45.6	45.3	35.5
SeCu (ICCV2023)	56.1	59.0	54.5	40.3	34.8	33.7	35.8	19.9	51.6	50.8	48.2	39.1
LFSS (ICML2025)	58.1	63.1	59.7	43.3	43.5	42.2	46.8	28.6	46.6	43.7	46.0	34.0
ConMix (ICLR2025)	61.6	65.4	59.8	45.6	42.8	41.0	43.9	27.7	48.9	49.7	**48.8**	**35.9**
MiniClustering (ICLR2026)	62.3	67.4	65.6	**50.4**	44.2	43.6	47.2	**31.3**	49.8	51.1	48.4	34.9
MiniClustering (SimCLR)	55.7	61.3	56.8	39.4	40.6	38.7	40.6	23.2	45.1	47.0	51.8	29.5
Ours	**68.0**	**75.5**	**69.8**	45.9	**46.6**	**43.7**	**47.9**	30.1	**52.3**	**51.5**	45.6	32.5

Table 4: Clustering performance in percentage (%) on three benchmark datasets (imb_ratio = 10) across various methods. Best results are in bold; second-best results are underlined.

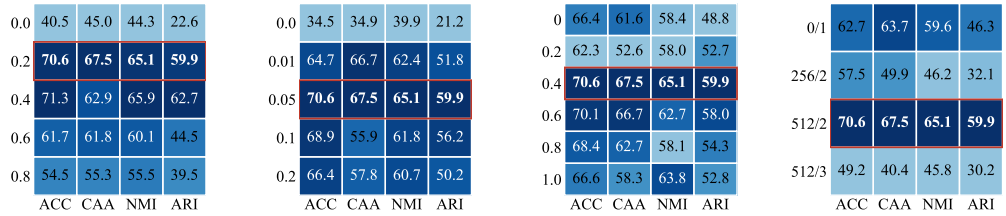
Method	CIFAR-10				CIFAR-20				STL-10			
	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI
SCAN (ECCV2020)	52.0	50.1	59.4	43.8	37.0	32.4	37.8	23.1	47.1	45.4	47.5	34.2
SDCLR (ICML2021)	38.9	44.3	42.5	26.5	37.8	35.9	39.6	22.9	34.6	37.9	32.2	17.3
CC (AAAI2021)	40.6	27.5	43.9	18.8	19.9	14.3	21.9	1.1	43.0	35.3	44.7	25.4
IDFD (ICLR2021)	47.5	54.9	48.4	33.1	28.7	27.2	28.6	15.1	38.6	34.7	36.8	22.8
CoNR (NeurIPS2023)	31.4	44.3	29.2	17.8	20.3	17.9	17.3	8.0	34.8	32.5	30.7	21.0
ProPos (TPAMI2022)	46.1	49.3	52.5	34.2	36.8	33.6	40.1	22.5	35.6	38.2	37.2	23.9
DMICC (AAAI2023)	36.6	39.5	36.8	25.9	24.7	21.9	20.7	10.1	41.3	41.9	38.7	30.3
SeCu (ICCV2023)	51.5	57.0	54.8	39.6	32.0	30.7	33.3	17.6	48.2	**49.6**	**50.0**	**36.8**
LFSS (ICML2025)	56.3	59.7	57.9	43.0	41.1	40.0	46.8	28.4	43.3	42.3	45.4	30.8
ConMix (ICLR2025)	53.3	58.2	57.1	40.8	41.7	39.3	43.6	27.0	47.4	48.7	48.2	33.9
MiniClustering (ICLR2026)	59.9	58.2	60.3	52.5	35.8	35.8	41.2	23.6	48.0	42.1	44.2	33.0
MiniClustering (SimCLR)	51.5	58.5	55.5	37.5	34.5	33.8	37.4	18.0	41.4	41.1	40.3	24.9
Ours	**70.6**	**67.5**	**65.1**	**59.9**	**48.3**	**44.2**	**48.2**	**31.6**	**50.8**	49.5	44.3	32.9

4.2 Main Results

Results on fine-grained long-tailed benchmarks. Table 1 reports clustering performance on CUB-200-2011, Stanford Dogs and Oxford Flowers; all methods are evaluated under the same frozen encoder. D4C outperforms the compared methods on ACC, CAA and NMI, trailing ConMix only on Oxford Flowers ARI by 2.1. Since these baselines all optimize a single final feature space(Uziel et al., 2025; Liu et al., 2024; Li et al., 2025b; Li & Jia, 2025; Li et al., 2025a), these results corroborate our central claim that fine-grained tail structures must be preserved across representation levels before final-layer rebalancing.

Generalization to larger-scale benchmarks. Table 2 extends D4C to Tiny-ImageNet-LT, ImageNet-100-LT and ImageNet-50-LT (ratio 10). SimCLR features are too weak for reliable clustering on these larger datasets, so we use DINOv3 weights(Siméoni et al., 2025). D4C outperforms all baselines on every metric, confirming that the coverage-risk decomposition generalizes beyond the CIFAR/STL regime.

Results on CIFAR-10, CIFAR-20 and STL-10. Tables 3 and 4 report clustering performance at imbalance ratios 5 and 10. MiniClustering(Li et al., 2025a), the strongest baseline, is re-run under the same public SimCLR(Chen et al., 2020) backbone as D4C. D4C achieves state-of-the-art results across almost all metrics and datasets, with the advantage growing as the imbalance worsens. Concretely, on CIFAR-10 the margin over the strongest baseline widens from 5.7 ACC points at ratio 5 to 10.7 points at ratio 10. The CAA gains are consistently large, indicating that layer-wise coverage correction preserves tail-class structures that final-layer rebalancing loses.



(a) β_{hid} (b) λ_{hid} (c) α (d) width / layers

Figure 2: Hyperparameter sensitivity of D4C on CIFAR-10 (imb_ratio = 10). Each heatmap reports clustering performance under different hyperparameter configurations. Entries that are both bolded and enclosed in a red box correspond to the default setting used in the main experiments.

4.3 Ablation Studies

Component ablation. Table 5 ablates each mechanism of D4C on CIFAR-10 (imb_ratio = 10). EMA teacher freezing stabilizes the plateau but cannot hold it through the full schedule once targets are frozen. For the w/o $\mathcal{L}_{\text{geo}}^{(l)}$ and w/o $\mathcal{L}_{\text{vmf}}^{(l)}$ variants, training collapses after the reported epoch, so their values are reported at the last stable epoch (Table 5); removing either term sharply degrades all metrics, confirming that global separation and local compactness must act jointly. Additional analyses, including the effect of α at a milder imbalance ratio and the supervised oracle upper bound(Van Gansbeke et al., 2020; Huang et al., 2022), are reported in the Appendix G and Appendix H, respectively. Quantitatively, the supervised oracle upper bound of 91.4 ACC (Appendix H) sits 20.8 points above the unsupervised model, indicating that most residual error on CIFAR-10 stems from self-generated Sinkhorn targets rather than from the frozen backbone.

Table 5: Component ablation on CIFAR-10 (imb_ratio = 10). For the w/o $\mathcal{L}_{\text{geo}}^{(l)}$ and w/o $\mathcal{L}_{\text{vmf}}^{(l)}$ variants, training becomes unstable after the reported epoch; we report the value at the last stable epoch.

Variant	ACC	CAA	NMI	ARI
Full D4C	70.6	67.5	65.1	59.9
w/o prototype anchoring ($\beta_{\text{hid}} = 0$)	40.5	45.0	44.3	22.6
w/o long-tail marginal ($\alpha = 0$)	66.4	61.6	58.4	48.8
w/o over-segmentation ($K = 10$)	52.3	53.2	60.1	43.7
w/o density-aware forward layers	62.7	63.7	59.6	46.3
w/o EMA teacher freezing	69.6	63.2	60.8	57.0
w/o $\mathcal{L}_{geo}^{(l)}$ only	32.7	33.2	32.6	13.4
w/o $\mathcal{L}_{vmf}^{(l)}$ only	38.2	36.2	39.9	18.9

Design choices. Figure 2 reports the sensitivity of the four key hyperparameters on CIFAR-10 under the imbalanced setting. Overall, D4C remains stable around the default configurations. In Fig. 2a, β_{hid} governs prototype anchoring; performance improves as anchoring becomes effective and then saturates, while overly strong anchoring sacrifices CAA. Fig. 2b shows that λ_{hid} is critical for hidden-layer coverage shaping; disabling it leads to a clear failure, and excessive values also hurt the balance among metrics. Fig. 2c indicates that α interpolates between statistical matching and long-tail balancing, with a moderate value giving the best overall trade-off. Fig. 2d suggests that forward-layer capacity and depth have a broad stable region, and increasing them beyond the default does not bring consistent gains. Complementary training-dynamics, representation-evolution, long-tail, and computational-cost analyses are in Appendices J–M.

5 Conclusions

In this work, we introduced **D4C**, a hierarchical coverage risk optimization framework that shapes sample coverage, center separation, and local density at every forward layer, rather than only in the final embedding. The PAC analysis assigns each objective a distinct role in the coverage-risk bound, and ablations confirm that all components are necessary. D4C attains state-of-the-art clustering across long-tailed and fine-grained benchmarks, with the largest gains in tail-class recall. A limitation is that D4C inherits the frozen backbone: absolute accuracy on fine-grained data remains low, and the coverage objective favors ACC/CAA, occasionally conceding NMI/ARI to single-stage methods on datasets such as STL-10 (Appendix I). Extending layer-wise coverage control to end-to-end representation learning and open-world category discovery are promising directions.

AI USE STATEMENT

We used generative AI tools for manuscript editing, language polishing, reference formatting, LaTeX assistance, and for revising and debugging code. We did not use generative AI for generating research ideas, developing theoretical models, deriving mathematical proofs, designing methods or experiments, or analyzing data. All AI-assisted work was reviewed and verified against our experimental records and cited sources. We take responsibility for the final content, including text, claims, and artifacts.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. The study uses only publicly available benchmark datasets and does not involve human subjects, crowdsourcing, personal data, privacy-sensitive information, or high-risk model releases. All external datasets and code assets are properly cited and used under their original licenses. The proposed method is a general unsupervised clustering framework. We do not foresee direct negative societal impacts or misuse risks that require special safeguards.

REPRODUCIBILITY STATEMENT

We have taken several measures to ensure the reproducibility of this work. The complete algorithmic details of D4C, including the hierarchical coverage risk optimization, optimal transport formulation, geometric repulsive potential, vMF density pulling, prototype anchoring, over-segmentation with Ward merging, and EMA teacher freezing, are provided in Sec. 3. The experimental protocols, dataset construction, backbone choices, and all hyperparameters are described in Sec. 4.1. Theoretical derivations and additional analyses, including the PAC upper bound of coverage risk, the monotonicity of the geometric kernel determinant, and the discussion on spherical covering and separation, are provided in Appendices A–E. The source code will be publicly released, and pseudocode is provided in the paper (Appendix E). All datasets used are publicly available benchmarks, and the long-tailed training sets are constructed by exponential sub-sampling with a fixed imbalance ratio, as detailed in Sec. 4.1.

REFERENCES

Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models. *arXiv preprint arXiv:2303.17896*, 2023.

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Károly Böröczky Jr. Finite packing and covering. *Cambridge tracts in mathematics*, 154, 2004.

Stephane Boucheron, Gabor Lugosi, and Pierre Massart. A non asymptotic theory of independence, 2013.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

John Horton Conway, Neil James Alexander Sloane, Etsuko Bannai, J Leech, SP Norton, AM Odlyzko, RA Parker, L Queen, and BB Venkov. *Sphere packings, lattices and groups*, volume 290. Springer New York, 1999.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22, 1977.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm, 1979.

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*, pp. 770–778, 2016.

Cuong Manh Hoang, Yeejin Lee, and Byeongkeun Kang. Unsupervised contrastive learning using out-of-distribution data for long-tailed dataset. *Neurocomputing*, 649:130779, 2025.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7509–7524, 2022.

Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9865–9874, 2019.

Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In *International conference on machine learning*, pp. 4927–4939. PMLR, 2021.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Yunji Kim and Jung-Woo Ha. Contrastive fine-grained class clustering via generative adversarial networks. *arXiv preprint arXiv:2112.14971*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8547–8555, 2021.

Zhixin Li and Yuheng Jia. Conmix: Contrastive mixup at representation level for long-tailed deep clustering. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhixin Li, Yuheng Jia, Guanliang Chen, Hui Liu, and Junhui Hou. Mini-cluster guided long-tailed deep clustering. In *The Fourteenth International Conference on Learning Representations*, 2025a.

Zhixin Li, Yuheng Jia, Hui LIU, and Junhui Hou. Learning from sample stability for deep clustering. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=gtVW8CO6Wf.

Honglin Liu, Peng Hu, Changqing Zhang, Yunfan Li, and Xi Peng. Interactive deep clustering via value mining. *Advances in Neural Information Processing Systems*, 37:42369–42387, 2024.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2537–2546, 2019.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1): 148–188, 1989.

James B McQueen. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th berkeley symposium on math. stat. and prob.*, pp. 281–297, 1967.

Ioannis Maniadis Metaxas, Georgios Tzimiropoulos, and Ioannis Patras. Divclust: Controlling diversity in deep clustering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3418–3428. IEEE, 2023.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Bo Peng, Yuanwei Hu, Bo Liu, Ling Chen, Jie Lu, and Zhen Fang. Delving into spectral clustering with vision-language representations. In *The Fourteenth International Conference on Learning Representations*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Liangliang Shi, Zhaoqi Shen, and Junchi Yan. Double-bounded optimal transport for advanced clustering and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14982–14990, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6483–6492. IEEE, 2019.

Yaling Tao, Kentaro Takagi, and Kouta Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=e12NDM7wkEY.

Roy Uziel, Irit Chelly, Oren Freifeld, and Ari Pakman. Clustering via self-supervised diffusion. *arXiv preprint arXiv:2507.04283*, 2025.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pp. 268–285. Springer, 2020.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Roman Vershynin. High-dimensional probability. *Cambridge Series in Statistical and Probabilistic Mathematics*, 47, 2019.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8927–8948, 2021.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pp. 478–487. PMLR, 2016.

Xulun Ye, Benyu Wu, Jie Hong, and Kun Zhou. Nonparametric deep fine-grained clustering with low-rank guided vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2434–2444, 2026.

Lingjie Yi, Jiachen Yao, Weimin Lyu, Haibin Ling, Raphael Douady, and Chao Chen. Geometry of long-tailed representation learning: Rebalancing features for skewed distributions. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chuyu Zhang, Hui Ren, and Xuming He. P^2ot: Progressive partial optimal transport for deep imbalanced clustering. *arXiv preprint arXiv:2401.09266*, 2024.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9719–9728, 2020.

Sheng Zhou, Hongjia Xu, Zhuonan Zheng, Jiawei Chen, Zhao Li, Jiajun Bu, Jia Wu, Xin Wang, Wenwu Zhu, and Martin Ester. A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. *ACM Computing Surveys*, 57(3):1–38, 2024.

Zhihan Zhou, Jiangchao Yao, Yan-Feng Wang, Bo Han, and Ya Zhang. Contrastive learning with boosted memorization. In *International Conference on Machine Learning*, pp. 27367–27377. PMLR, 2022.

CONTENTS

A PAC Characterization of Coverage Risk **16**

B Adaptive Coverage Bound for Data-Dependent Centers **16**

C The vMF Posterior as Entropic Optimal Transport **17**

D Monotonicity in the Equiangular Case for Geometric Kernel Determinant **18**

E Supplementary Notes on Spherical Covering and Separation **18**

F Training Algorithm **19**

G Effect of α at a Milder Imbalance Ratio **20**

H Supervised Oracle Upper Bound **21**

I Per-Dataset Discussion **21**

J Computational Cost **22**

K Representation Visualization **22**

L Training Dynamics on CIFAR-10 **23**

M Long-Tail Behavior and Prototype Merging **25**

A PAC CHARACTERIZATION OF COVERAGE RISK

Let the features of a given layer be random vectors $x \sim P$ on the unit sphere $\mathbb{S}^{d-1}$, and the center set be $C = \{c_1, \ldots, c_M\} \subset \mathbb{S}^{d-1}$. For a fixed radius $\tau > 0$, the coverage risk (Vapnik, 1999) is defined as

$$R_\tau(C) = \Pr_{x \sim P}\Big[\min_m \|x - c_m\| > \tau\Big].$$

Define the random variable

$$Y = \min_m \|x - c_m\|^2.$$

Since $x, c_m \in \mathbb{S}^{d-1}$, we have $\|x - c_m\|^2 = 2 - 2\langle x, c_m\rangle \in [0, 4]$, and thus $Y \in [0, 4]$. By Markov's inequality (Boucheron et al., 2013),

$$R_\tau(C) = \Pr[Y > \tau^2] \le \frac{\mathbb{E}[Y]}{\tau^2}.$$

Let $x_1, \ldots, x_N$ be independent and identically distributed samples. $Y_i = \min_m \|x_i - c_m\|^2$ is a bounded random variable taking values in $[0, 4]$. The one-sided Hoeffding's inequality (Hoeffding, 1963) states that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\mathbb{E}[Y] \le \frac{1}{N}\sum_{i=1}^{N} Y_i + 4\sqrt{\frac{\ln(1/\delta)}{2N}}.$$

Substituting this into the Markov upper bound yields

$$R_\tau(C) \le \frac{1}{\tau^2}\left(\frac{1}{N}\sum_{i=1}^{N} \min_m \|x_i - c_m\|^2 + 4\sqrt{\frac{\ln(1/\delta)}{2N}}\right).$$

Defining the empirical mean nearest-peak cost as

$$\Psi(C) = \frac{1}{N}\sum_{i=1}^{N} \min_m \|x_i - c_m\|^2,$$

the above inequality recovers Eq. 10 in the main text:

$$R_\tau(C) \le \frac{1}{\tau^2}\left(\Psi(C) + 4\sqrt{\frac{\ln(1/\delta)}{2N}}\right).$$

This upper bound depends neither on the feature dimension d nor on the number of peaks M, but solely on the bounded range $[0, 4]$ of Y.

The derivation above treats the center set C as fixed and independent of the sample. Since D4C estimates its centers (density peaks or prototypes) from the same data, a uniform bound over all candidate center sets is required; Appendix B derives this adaptive bound, which adds an $O(\sqrt{dM/N})$ complexity penalty to the $O(N^{-1/2})$ term while keeping the same optimizable leading term $\Psi(\hat{C})$.

B ADAPTIVE COVERAGE BOUND FOR DATA-DEPENDENT CENTERS

The bound in Appendix A treats the center set C as fixed and independent of the sample $S = \{x_1, \ldots, x_N\}$. In D4C the centers (density peaks or prototypes) are estimated from the same data, so we now give a uniform bound over all candidate center sets.

Let $\mathcal{C}_M = (S^{d-1})^M$ denote all ordered M-tuples of unit vectors, and define the function class

$$\mathcal{F}_M = \big\{f_C : x \mapsto \min_{m \le M} \|x - c_m\|^2 \;:\; C = (c_1, \ldots, c_M) \in \mathcal{C}_M\big\},$$

with $0 \le f_C \le 4$. Each function is 2-Lipschitz in C: for any $C, C' \in \mathcal{C}_M$,

$$|f_C(x) - f_{C'}(x)| \le \max_m \big|\|x - c_m\|^2 - \|x - c'_m\|^2\big| = 2\max_m |\langle x, c_m - c'_m\rangle| \le 2\max_m \|c_m - c'_m\|.$$

Consequently an $\frac{\varepsilon}{2}$-cover of S^{d-1} induces an ε-cover of $\mathcal{F}_M$ in $\|\cdot\|_\infty$, and the covering number of the sphere obeys $\mathcal{N}(S^{d-1}, \varepsilon, \|\cdot\|) \le (1 + 4/\varepsilon)^d$ (Vershynin, 2019; Wainwright, 2019). Hence

$$\ln \mathcal{N}(\mathcal{F}_M, \varepsilon, \|\cdot\|_\infty) \le dM \ln\big(1 + \tfrac{8}{\varepsilon}\big). \tag{14}$$

By a standard symmetrization argument (Bartlett & Mendelson, 2002) followed by McDiarmid's inequality (McDiarmid et al., 1989) and Dudley chaining (Boucheron et al., 2013; Wainwright, 2019), the worst-case gap between the population mean and the empirical mean over $\mathcal{F}_M$ satisfies, with probability at least $1 - \delta$,

$$\Delta_N(\delta) := \sup_{C \in \mathcal{C}_M} \Big| \mathbb{E}_{x\sim P}[f_C(x)] - \frac{1}{N}\sum_{i=1}^{N} f_C(x_i) \Big| \le 2\,\mathcal{R}_N(\mathcal{F}_M) + 4\sqrt{\tfrac{2\ln(1/\delta)}{N}}, \tag{15}$$

where $\mathcal{R}_N$ is the empirical Rademacher complexity. Inserting the covering estimate 14 into the chaining integral gives

$$\mathcal{R}_N(\mathcal{F}_M) \le 12 \int_0^4 \sqrt{\tfrac{\ln \mathcal{N}(\mathcal{F}_M, t, \|\cdot\|_\infty)}{N}}\, \mathrm{d}t \le 12\sqrt{\tfrac{dM}{N}} \int_0^4 \sqrt{\ln\big(1 + \tfrac{8}{t}\big)}\, \mathrm{d}t = C_0 \sqrt{\tfrac{dM}{N}},$$

where $C_0 := 12 \int_0^4 \sqrt{\ln(1 + 8/t)}\, \mathrm{d}t < \infty$ is an absolute constant.

Combining the above with Markov's inequality, for any (possibly data-dependent) center set $\hat{C} \in \mathcal{C}_M$,

$$R_\tau(\hat{C}) = \Pr_{x\sim P}\big[f_{\hat{C}}(x) > \tau^2\big] \le \frac{\mathbb{E}_{x\sim P}[f_{\hat{C}}(x)]}{\tau^2} \le \frac{1}{\tau^2}\Big(\Psi(\hat{C}) + \Delta_N(\delta)\Big). \tag{16}$$

With probability at least $1 - \delta$ over the training sample, this yields the adaptive coverage bound

$$R_\tau(\hat{C}) \le \frac{1}{\tau^2}\Big(\Psi(\hat{C}) + 2C_0\sqrt{\tfrac{dM}{N}} + 4\sqrt{\tfrac{2\ln(1/\delta)}{N}}\Big). \tag{17}$$

The adaptive bound 17 has the same structure as the fixed-center bound 10: the empirical nearest-center cost $\Psi(\hat{C})$ remains the dominant, optimizable term, while the complexity penalty now grows with the ambient dimension d and the number of centers M. Consequently, minimizing $\Psi(\hat{C})$ through the OT and vMF objectives of Secs. 3.1–3.2, while keeping centers non-degenerate through the geometric repulsion of Eq. 4, optimizes the leading term of a valid finite-sample coverage guarantee for data-dependent centers.

C The vMF Posterior as Entropic Optimal Transport

Proposition. At the EM fixed point (Dempster et al., 1977) of the vMF mixture (Banerjee et al., 2005) (Eqs. 5–6), the posterior satisfies $\sum_i q_{im}^{(l)} = N\omega_m^{(l)}$, and $q_{im}^{(l)}/N$ is the unique entropy-regularized optimal transport plan (Cuturi, 2013; Gabriel & Marco, 2019),

$$\arg\min_{\Pi \ge 0} \sum_{i,m} \Pi_{im}\big(-\kappa_m^{(l)} \langle f_i^{(l)}, \mu_m^{(l)} \rangle\big) + \sum_{i,m} \Pi_{im}\big(\log \Pi_{im} - 1\big) \quad \text{s.t.} \quad \Pi\, \mathbf{1}_M = \tfrac{1}{N}\mathbf{1}_N, \;\; \Pi^\top \mathbf{1}_N = \omega^{(l)},$$

with regularization strength $\varepsilon = 1$; the vMF normalizers $C_d(\kappa_m^{(l)})$ are absorbed into the column potentials of the entropic plan.

Proof. With $\varepsilon = 1$, the KKT conditions of the entropic OT problem give $\Pi_{im} = \beta_m\, e^{\kappa_m^{(l)} \langle f_i^{(l)}, \mu_m^{(l)} \rangle} / Z_i$, where β_m is the column Lagrange multiplier and Z_i normalizes row i to $1/N$; identifying $\beta_m \propto \omega_m^{(l)} C_d(\kappa_m^{(l)})$ reproduces Eq. 6. The column constraint $\sum_i \Pi_{im} = \omega_m^{(l)}$ is exactly the EM mass update $\omega_m^{(l)} = \frac{1}{N}\sum_i q_{im}^{(l)}$, which holds at the EM fixed point. Strict convexity of the entropic objective gives uniqueness. On the sphere, $-\kappa_m \langle f, \mu_m \rangle$ differs from $\frac{\kappa_m}{2}\|f - \mu_m\|^2$ by a constant, so the plan is the entropic counterpart of the squared-distance OT in Sec. 3.1.

D Monotonicity in the Equiangular Case for Geometric Kernel Determinant

The geometric repulsive potential (Kulesza & Taskar, 2012) employs a kernel matrix based on Gaussian similarity:

$$S_{ij} = \exp\left(-\frac{\|c_i - c_j\|^2}{2\sigma^2}\right).$$

When the center set approaches the maximally separated configuration, all off-diagonal elements S_{ij} tend to zero, and the determinant approaches 1. Conversely, if centers are close to each other, $S_{ij} \to 1$, the kernel matrix approaches the all-ones matrix, its eigenvalues are M once and 0 with multiplicity $M-1$. Therefore, maximizing $\det(S)$ favors enlarging inter-center distances.

To provide a computable evidence, consider the equiangular case: all off-diagonal elements $S_{ij} = \rho$ ($0 \le \rho < 1$), and diagonal elements are 1. Then

$$S_\rho = (1-\rho)I_M + \rho \mathbf{1}\mathbf{1}^\top.$$

Its eigenvalues are $1 + (M-1)\rho$ (multiplicity 1) and $1-\rho$ (multiplicity $M-1$), hence

$$\det(S_\rho) = (1 + (M-1)\rho)(1-\rho)^{M-1}.$$

Differentiating with respect to ρ:

$$\frac{d}{d\rho} \ln \det(S_\rho) = \frac{M-1}{1+(M-1)\rho} - \frac{M-1}{1-\rho} = -\frac{M(M-1)\rho}{(1+(M-1)\rho)(1-\rho)} \le 0.$$

This is strictly negative for $\rho > 0$. Thus $\det(S_\rho)$ decreases monotonically as ρ increases; maximizing the determinant is equivalent to minimizing ρ, which means maximizing the inter-center distance. This monotonicity provides concrete analytical evidence that the log-det repulsive force is a surrogate objective for geometric dispersion.

For the general non-equiangular case, Hadamard's inequality (Horn & Johnson, 2012) alone only yields the asymptotic statement $\det(S) \to 1 \Leftrightarrow S \to I_M$. The following bound upgrades it to a quantitative, per-pair separation guarantee.

Proposition (separation lower bound). Let $d_{\min} = \min_{i \ne j} \|c_i - c_j\|$ be the minimum inter-center distance. For every configuration with $0 < \det(S) \le 1$,

$$d_{\min}^2 \;\ge\; \sigma^2 \ln \frac{1}{1 - \det(S)},$$

and the right-hand side is strictly increasing in $\det(S)$. Hence maximizing $\det(S)$—equivalently minimizing $\mathcal{L}_{\text{geo}}^{(l)}$ in Eq. 4—provably enlarges the guaranteed pairwise separation of every center pair, in arbitrary configurations.

Proof. For any $i \ne j$, the principal submatrix on $\{i,j\}$ is $\left(\begin{smallmatrix} 1 & S_{ij} \\ S_{ij} & 1 \end{smallmatrix}\right)$ with determinant $1 - S_{ij}^2$. Since S is positive semidefinite, Fischer's inequality (Horn & Johnson, 2012) gives $\det(S) \le (1 - S_{ij}^2)\,\det(S_{[M]\setminus\{i,j\}})$, and Hadamard's inequality bounds the complementary block by $\prod_{k \notin \{i,j\}} S_{kk} = 1$. Thus $S_{ij} \le \sqrt{1 - \det(S)}$, and inserting $S_{ij} = \exp(-\|c_i - c_j\|^2/(2\sigma^2))$ and taking logarithms yields the claim. The bound is tight for an antipodal pair ($M = 2$, where $\det(S) = 1 - e^{-4/\sigma^2}$ and $d_{\min}^2 = 4$). With the γI jitter used in Eq. 4, the same argument gives $\det(S + \gamma I) \le ((1+\gamma)^2 - S_{ij}^2)(1+\gamma)^{M-2}$; the monotonicity in $\det(S + \gamma I)$ is unchanged.

E Supplementary Notes on Spherical Covering and Separation

The covering problem for M points on the unit sphere (finding a point set that minimizes the covering radius $\max_x \min_m \|x - c_m\|$) and the separation problem (finding a point set that maximizes the minimum pairwise distance $\min_{i \ne j} \|c_i - c_j\|$) are classical geometric problems (Conway et al., 1999; Böröczky Jr, 2004). In high dimensions, the two are not strictly equivalent: there exist configurations with high separation but still large covering radius (e.g., an equatorial ring configuration

lacking coverage near the poles). Thus, one cannot simply assert that maximizing the determinant is equivalent to minimizing the covering radius.

Our theoretical framework adopts the following conservative statements to ensure rigor:

1. The geometric repulsive potential maximizes $\det(S)$, which increases the minimum inter-peak distance by suppressing off-diagonal elements of the kernel matrix;
2. A larger minimum inter-peak distance helps prevent multiple peaks from falling into the same high-density region, thereby avoiding covering holes caused by peak redundancy;
3. Empirically, greater inter-peak separation correlates with a lower covering radius, but we do not claim a strictly monotonic or equivalent relationship between the two.

Separation transfer. Take $m(k)$ to be the maximum-weight one-to-one matching (Kuhn, 1955) between prototypes and peaks (defined whenever at least K peaks survive pruning), and suppose the anchoring gap obeys $\|w_k - \mu_{m(k)}\| \le \zeta$ for all k. Then for every $k \neq j$,

$$\|w_k - w_j\| \;\ge\; \|\mu_{m(k)} - \mu_{m(j)}\| - 2\zeta.$$

Eq. 9 maximizes $\langle \mu_m, w_k \rangle = \frac{1}{2}(2 - \|\mu_m - w_k\|^2)$, i.e., it minimizes exactly the gaps entering ζ. Hence the peak separation enforced by $\mathcal{L}_{\text{geo}}^{(l)}$ (Appendix D) propagates to the output-layer center set $\{w_k\}$ up to 2ζ, keeping the worst-case radius of Eq. 11 non-degenerate.

The statement in the main text that the geometric repulsive potential suppresses center redundancy should be understood as a biased regularization trend induced by geometric separation, rather than a conclusion strictly derived from the aforementioned lemmas. This trend complements the direct minimization of the mean cost $\Psi(C)$ by optimal transport: optimal transport pulls centers toward centroids, while the geometric repulsive potential prevents centers from degenerating and coinciding; together, they form a viable surrogate for coverage risk.

F Training Algorithm

The complete D4C procedure is summarized in Algorithm 1. It instantiates the layer-wise coverage-risk objectives of Secs. 3.1–3.3 within the stabilization scheme of Sec. 3.4. Here $\tilde{f}_i^{(l)}$ denotes the layer-l hidden representation ($l = 1, \ldots, L-1$) of the second, stop-gradient view, computed through the same h_θ; $\mathcal{L}_{\text{sk}}$ uses the last one, $\tilde{f}_i^{(L-1)}$.

Loss definition. The output-layer loss $\mathcal{L}_{\text{sk}}$ is a cross-entropy between the student predictions and the Sinkhorn targets of the EMA teacher:

$$p_{ik} = \frac{\exp\langle f_i^{(L-1)}, w_k \rangle}{\sum_{k'} \exp\langle f_i^{(L-1)}, w_{k'} \rangle}, \qquad t_{ik} = \text{Sinkhorn}_\varepsilon\big(-\langle \tilde{f}_i^{(L-1)}, \bar{w}_k \rangle;\, a = \tfrac{1}{N}\mathbf{1}_N,\, b = m\big) \tag{18}$$

$$m_k = \frac{\pi_k^\alpha}{\sum_{k'} \pi_{k'}^\alpha}, \qquad \mathcal{L}_{\text{sk}} = -\frac{1}{N} \sum_{i,k} t_{ik} \log p_{ik}. \tag{19}$$

Predictions use the student prototypes w_k; targets use the frozen teacher prototypes $\bar{w}_k$, with $\varepsilon = 0.5$ and 3 Sinkhorn iterations.

Implementation notes.

- vMF-EM uses K-means initialization and 30 iterations with $\kappa \in [1, 500]$.
- The vMF pull uses a detached per-batch posterior at temperature 0.15; the EM concentration κ_m is absorbed into this posterior, so the pull weight in Eq. 8 is 1.
- The Sinkhorn target uses ε=0.5 with 3 iterations, and the EMA teacher decay is ρ_e=0.995; warm-up lasts T_w=10 epochs.
- Adam is used with learning rate 5×10^{-3} and gradient clipping at max norm 5.0; Ward merging, the final assignment and all reported metrics use the epoch-110 checkpoint inside the stable plateau (Appendix L).

- The vMF peak count is initialized as $M^{(l)} = \min(M_{\max}, \max(2K, N/200))$ with $M_{\max} = 200$, where $K = \rho K^*$; low-mass peaks are pruned thereafter, so the surviving peak count is data-dependent.
- The geometric kernel bandwidth is a learned parameter initialized at $\sigma = 1.0$ and clamped to $[0.1, 5.0]$; the anchoring temperature is $\tau_p = 0.15$.
- **Fine-grained settings:** On Stanford Dogs, Oxford Flowers and CUB-200-2011, we widen the hidden layer to 1024 and set $\beta_{hid} = 0.4$, $\xi = 0.5$; other hyperparameters follow Sec. 4.1.
- All reported results are from a single run with seed 42 and deterministic cuDNN.

Algorithm 1 D4C: layer-wise coverage risk minimization

Require: unlabeled data $\{x_i\}_{i=1}^N$; over-segmentation factor ρ; warm-up epochs T_w; total epochs T
Require: hyperparameters $\alpha, \lambda_{\text{hid}}, \beta_{\text{hid}}, \xi, \rho_e$
Ensure: forward layers h_θ; merged cluster assignment $\hat{C}$

freeze encoder; initialize forward layers h_θ
warm up h_θ toward FCM targets for T_w epochs, with FCM running on the $K = \rho K^*$ prototypes; K^* is used only at evaluation
initialize prototypes $W \in \mathbb{R}^{H \times K}$ by K-means; $\bar{W} \leftarrow W$
fit vMF peaks on the hidden representations $\{f_i^{(l)}\}$ for each layer l; compute $q_{im}^{(l)}, \tilde{\pi}_m^{(l)}$ ▷ Eqs. 5–7
for $t = T_w{+}1$ **to** T **do**
  if $t \bmod 5 = 0$ **and** $t < T_w + \xi(T{-}T_w)$ **then**
    refit vMF peaks; update $q_{im}^{(l)}, \tilde{\pi}_m^{(l)}$ ▷ Eqs. 5–7
  end if
  for each batch with two views $(x_i^{(1)}, x_i^{(2)})$ **do**
    if $t \le T_w + \xi(T{-}T_w)$ **then**
      $\bar{W} \leftarrow \rho_e \bar{W} + (1{-}\rho_e) W$ ▷ EMA teacher, Sec. 3.4
    end if
    grad view $x_i^{(1)}$: compute $\{f_i^{(l)}\}_{l=1}^{L-1}$ ▷ Eq. 1
    no-grad view $x_i^{(2)}$: compute $\{\tilde{f}_i^{(l)}\}_{l=1}^{L-1}$; $t_{ik} \leftarrow \text{Sinkhorn}_\varepsilon\big(-\langle \tilde{f}_i^{(L-1)}, \bar{w}_k \rangle\big)$ with $m_k \propto \pi_k^\alpha$ ▷ Sec. 3.1
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{sk}} + \lambda_{\text{hid}} \sum_{l=1}^{L-1} \big(\mathcal{L}_{\text{geo}}^{(l)} + \mathcal{L}_{\text{vmf}}^{(l)}\big) + \beta_{\text{hid}} \sum_{l=1}^{L-1} \mathcal{L}_{\text{proto}}^{(l)}$ ▷ Eq. 13
    Adam step on h_θ with gradient clipping
  end for
end for
Ward-merge the EMA-averaged prototypes into K^* groups (K^* used for evaluation only) ▷ Eq. 12; assign each sample to its nearest group

G Effect of α at a Milder Imbalance Ratio

Table 6 reports the effect of the long-tail-aware marginal exponent α on CIFAR-10 with imbalance ratio 5. In contrast to the severe ratio-10 setting (Fig. 2c), a larger $\alpha = 0.8$ improves ACC and ARI while NMI drops by 4.6 compared with the default $\alpha = 0.4$, and CAA stays at a similar level.

Table 6: Impact of α on CIFAR-10 (imb_ratio = 5).

α	ACC	CAA	NMI	ARI
0.4	68.0	75.5	69.8	45.9
0.8	72.2	74.6	65.2	56.0

A plausible explanation is threefold. First, at a mild imbalance ratio, tail classes retain more samples—in CIFAR-10 the smallest class keeps 1,000 images at ratio 5 versus only 500 at ratio 10—so the FCM-estimated class prior is substantially more reliable: tail-class mass estimates suffer less sampling variance and tail clusters are less likely to be missed or merged during warm-up,

and a larger α aligns the transport marginal more faithfully with the true class support at little risk. Second, the balancing force of the uniform marginal is mainly needed to protect severely under-supported tail classes; at ratio 5 this protection is less critical, so the reduction in marginal mismatch dominates. Concretely, the largest improvement appears in ARI ($+10.1$), indicating that matched marginals stop head prototypes from absorbing tail samples and yield purer partitions, whereas CAA drops by only 0.9—in sharp contrast to ratio 10, where moving α from 0.4 to 1.0 costs 9.2 CAA (Fig. 2c). Third, because the prior is blended with the uniform one, $\alpha = 0.8$ at ratio 5 induces only a mild mass redistribution, which the over-segmented prototypes can absorb without destabilizing tail coverage; under severe imbalance, tail priors are small and noisy, and over-trusting them degrades CAA markedly. The optimal α thus increases as the imbalance becomes milder, consistent with the rationale of Sec. 3.1 that α interpolates between statistical matching and balancing. We retain $\alpha = 0.4$ as a single fixed default across benchmarks for robustness; per-benchmark tuning (e.g., $\alpha = 0.8$ at ratio 5) could yield further gains.

H Supervised Oracle Upper Bound

Table 7: Supervised oracle upper bound on CIFAR-10. "Unsupervised" is the full D4C with self-generated Sinkhorn targets; "Oracle" replaces them with ground-truth assignments.

Variant	ACC	CAA	NMI	ARI
Unsupervised (Full D4C)	70.6	67.5	65.1	59.9
Oracle targets	91.4	91.9	83.7	84.7

Following common practice for supervised upper-bound analysis in deep clustering (Van Gansbeke et al., 2020; Huang et al., 2022), we run a supervised oracle on CIFAR-10: all components remain unchanged, except that the self-generated Sinkhorn assignment targets are replaced by ground-truth labels. Table 7 reports the result, which serves as the performance upper bound of the framework under this backbone. Notably, because exact targets remove the long-tailed class-prior estimation altogether, the oracle upper bound is insensitive to the imbalance ratio; this is consistent with Appendix G, where the benefit of a larger α stems entirely from the reliability of the unsupervised class prior, which degrades as the imbalance worsens.

I Per-Dataset Discussion

Small-scale benchmarks. On CIFAR-20 the gain over the strongest baseline widens from 2.4 ACC points at ratio 5 to 12.5 at ratio 10—sharper than on CIFAR-10—because super-class labels blur class boundaries in the feature space, and these weak boundaries are exactly what final-layer rebalancing collapses; absolute values stay low since the super-classes group visually dissimilar CIFAR-100 classes. On STL-10, D4C wins ACC and, at ratio 5, CAA, but concedes NMI/ARI to SeCu, which we attribute to prototype over-segmentation: with only 13,000 images and broad, multi-modal classes, the 20 prototypes capture intra-class modes rather than the class partition, so Ward merging yields a less pure partition even when per-sample coverage is better.

Fine-grained benchmarks. Absolute accuracy (e.g., on CUB) remains low for all methods: the frozen MoCo v3 (Chen et al., 2021) backbone provides weak fine-grained features, and with up to 200 classes at ratio 10 the tail classes can retain only a few images, so the FCM prior and the vMF peaks are estimated from very few samples; the same over-segmentation effect concedes ARI to ConMix on Oxford Flowers. D4C improves the ranking rather than closing the gap.

Large-scale benchmarks. The most consistent gains occur on DINOv3 (Siméoni et al., 2025) features, where tail and fine-grained structure survives in $f^{(0)}$ so that layer-wise coverage correction refines genuine geometry rather than noise; on weaker SimCLR (Chen et al., 2020) features much of this structure is lost before the forward stack. The smaller ImageNet-50 margin is consistent: its 50 classes are already well separated in the DINOv3 space, leaving less room for correction. This matches the supervised oracle (Appendix H): with strong features, the self-generated targets become the main bottleneck, and coverage-risk minimization directly improves them.

J Computational Cost

All experiments run on a single NVIDIA RTX 4090 (24 GB). Table 8 reports the wall-clock time of each training phase on CIFAR-10. Feature caching and warm-up, which operate on cached features with the backbone offloaded to CPU, are inexpensive; post-warmup training dominates the budget. Figure 3 shows the peak GPU memory per epoch: memory stays far below the 24 GB limit throughout, and the warm-up phase has a negligible GPU footprint.

Table 8: Wall-clock time per training phase on CIFAR-10 (200 epochs). Feature cache covers frozen-backbone feature extraction (load or extract); initialization covers K-means (McQueen, 1967) and fuzzy C-means for projector warm-up. Evaluation and visualization snapshots are excluded.

Phase	Time (min)
Feature cache	0.3
K-means + FCM init	0.8
Warm-up (T_w=10 epochs)	0.1
Post-warmup (190 epochs)	53.8
Total	55.0

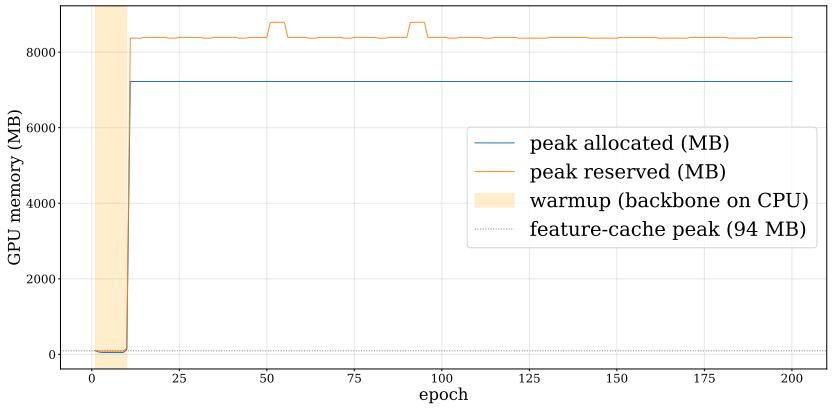


Figure 3: Peak GPU memory per epoch on CIFAR-10. Warm-up epochs run on cached features with the backbone offloaded to CPU; the dotted line marks the peak of the feature-caching phase.

K Representation Visualization

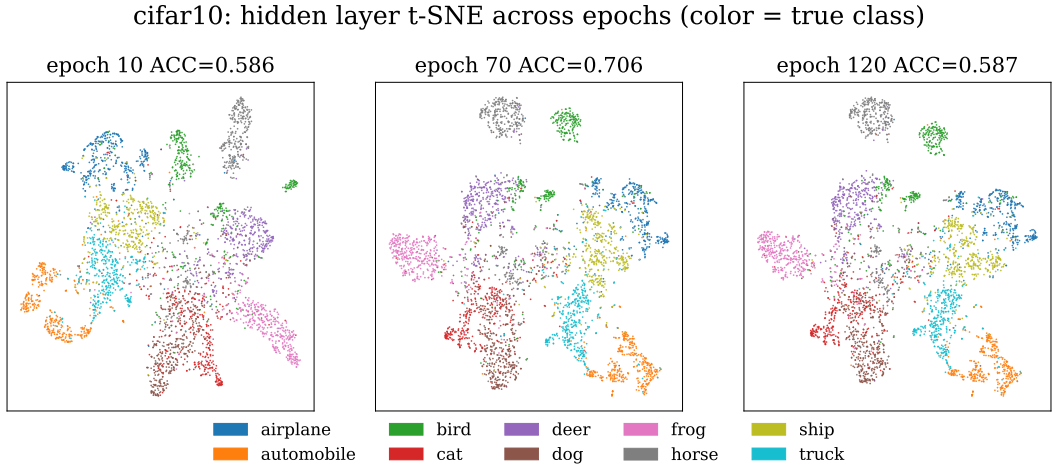


Figure 4: Hidden-layer t-SNE across training on CIFAR-10 (imb_ratio = 10), on the same stratified subsample. Left to right: epoch 10, epoch 70, epoch 120. Colors denote true classes.

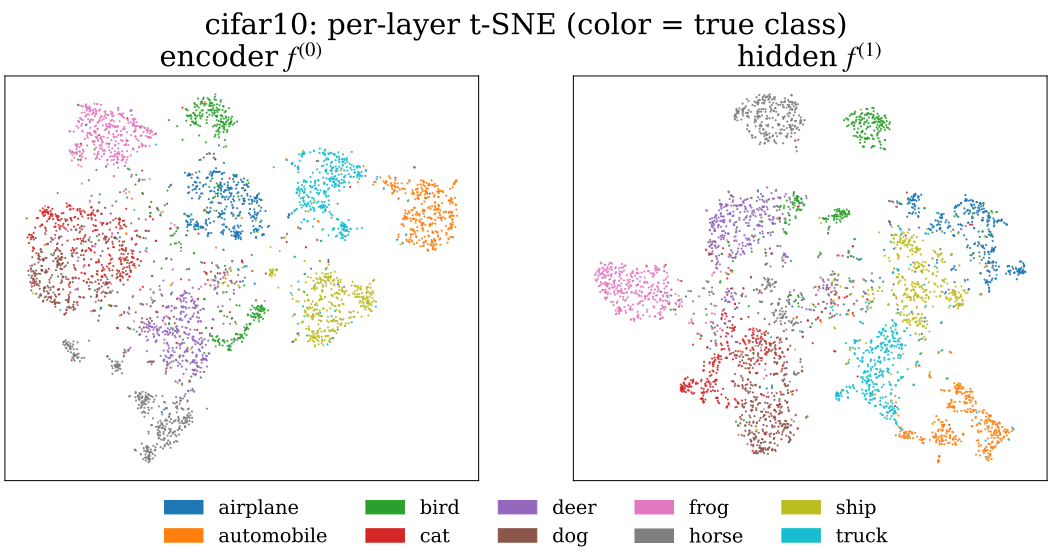


Figure 5: Per-layer t-SNE of D4C on CIFAR-10 (imb_ratio = 10). Left: frozen encoder output $f^{(0)}$; right: trained hidden representation $f^{(1)}$ (Eq. 1 with $L = 2$). Colors denote true classes.

Figure 4 visualizes the same stratified subsample of the hidden representation across three snapshots: epoch 10 (end of warm-up), the epoch-70 accuracy peak, and epoch 120 (post-peak, just before the SK-teacher freezing point $\xi = 0.6$). The ten classes are entangled at the end of warm-up and become clearly separated at the peak; after the peak, accuracy decays slowly as tail clusters gradually re-merge, so Table 4 reports the epoch-110 peak checkpoint, and the static snapshots of Figs. 5, 8 and 9 are drawn from the same checkpoint. The coverage-geometry curves of Fig. 7 improve monotonically through the same window, consistent with coverage risk being necessary but not sufficient for accuracy.

Figure 5 compares t-SNE (Van der Maaten & Hinton, 2008) embeddings of the frozen encoder output $f^{(0)}$ and the trained hidden representation $f^{(1)}$ of D4C. The hidden layer yields visibly tighter and better-separated clusters than the frozen backbone, particularly for tail classes, confirming that layer-wise coverage control refines, rather than merely re-uses, the backbone geometry.

L Training Dynamics on CIFAR-10

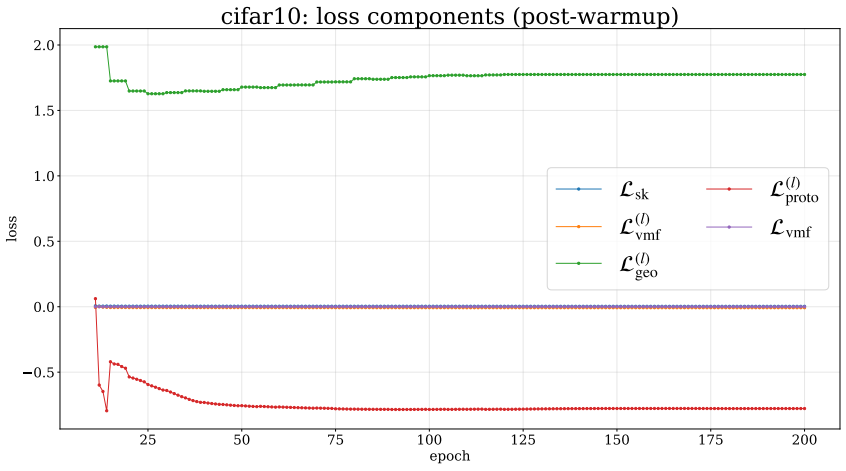


Figure 6: Loss decomposition over post-warmup training on CIFAR-10 (imb_ratio = 10). $\mathcal{L}_{\mathrm{sk}}$ is the final-layer risk term; the hidden-layer terms $\mathcal{L}^{(l)}_{\mathrm{geo}}$, $\mathcal{L}^{(l)}_{\mathrm{vmf}}$ and $\mathcal{L}^{(l)}_{\mathrm{proto}}$ are negative by design (Eqs. 4, 8, 9).

Figure 6 decomposes the joint objective (Eq. 13). $\mathcal{L}_{\text{sk}}$ is the dominant positive risk term of the final layer; the hidden-layer terms are negative by construction, since they maximize a log-determinant and cosine similarities (Eqs. 4, 8, 9). The long flat regime after epoch ≈120 is expected rather than a sign of poor optimization: the SK teacher freezes at $\xi = 0.6$ (Sec. 3.4), the vMF peaks are no longer re-estimated after the freezing epoch, and all loss weights are constant once the post-warmup ramp ends, so the targets defining these losses stop changing. A flat $\mathcal{L}_{\text{sk}}$ therefore certifies only alignment with the frozen teacher; to avoid any post-freezing drift, all reported results are taken from the stable window (Appendix F).

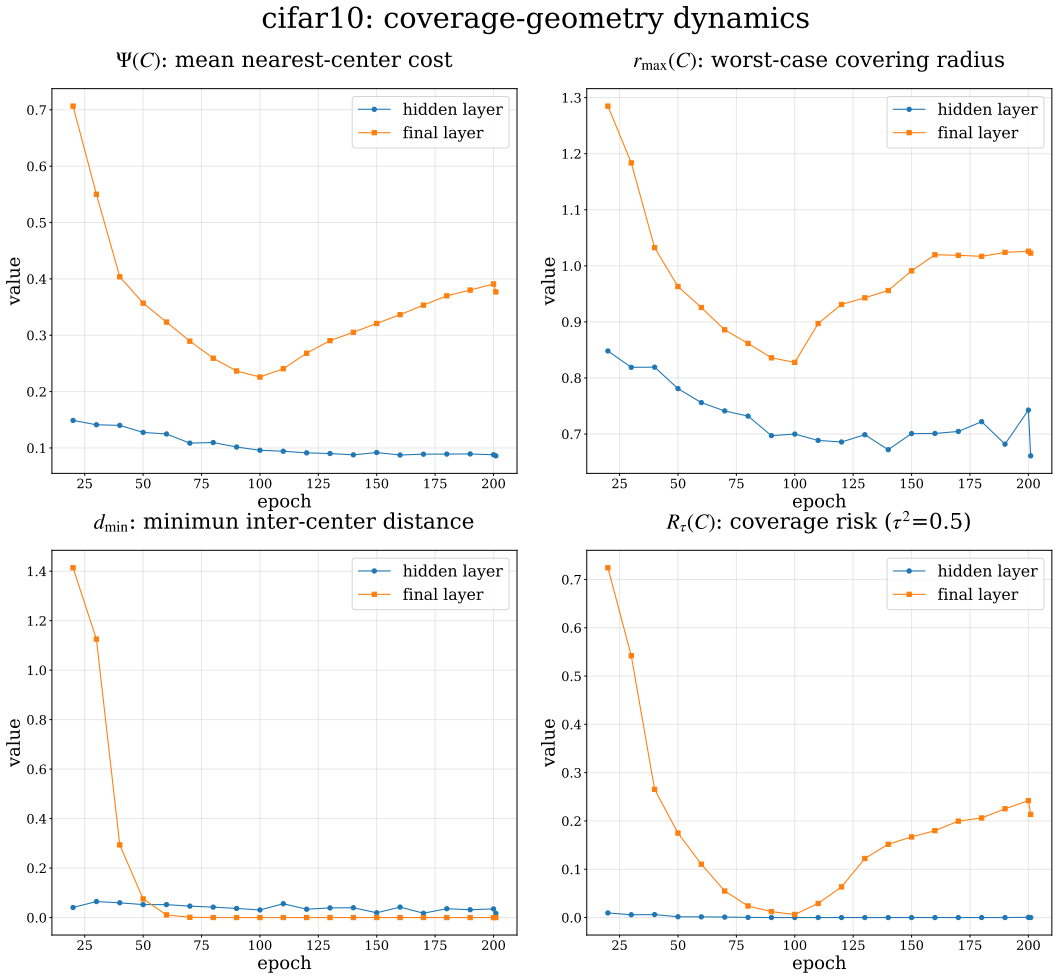


Figure 7: Coverage-geometry dynamics of D4C on CIFAR-10 (imb_ratio = 10). From top-left to bottom-right: the mean nearest-center cost $\Psi(C)$, the worst-case covering radius $r_{\max}(C)$, the minimum inter-center distance $d_{\min}$, and the coverage risk $R_\tau(C)$ with $\tau^2 = 0.5$ (Eqs. 10–11). Circle markers denote the hidden layer $f^{(1)}$; square markers denote the final layer. Post-warmup epochs only.

Figure 7 tracks the coverage geometry of the full D4C model over the post-warmup phase on CIFAR-10 (imb_ratio = 10, SimCLR backbone). For both the hidden and the final layer, $\Psi(C)$, $r_{\max}(C)$ and $R_\tau(C)$ decrease throughout training, while the minimum inter-center distance $d_{\min}$ stays bounded away from zero but decreases mildly before stabilizing: the OT and vMF objectives pull centers toward the data, and geometric repulsion (Eq. 4) prevents collapse, so coverage improves at a bounded cost in separation—the covering/separation trade-off of Appendix E. Notably, the coverage metrics keep improving through and beyond the accuracy plateau: ACC peaks at epoch 70 and then random-walks near the peak until roughly epoch 120, and Table 4 reports the epoch-110 checkpoint inside this window. The continued improvement of the coverage curves while clustering accuracy saturates

confirms that coverage risk is a necessary but not sufficient surrogate for clustering quality, which is why we report the epoch-110 checkpoint inside the stable window (Appendix F). Since the w/o $\mathcal{L}_{\text{geo}}^{(l)}$ and w/o $\mathcal{L}_{\text{vmf}}^{(l)}$ ablations collapse (Table 5), the monotone improvement here reflects the joint action of the two objectives.

M Long-Tail Behavior and Prototype Merging

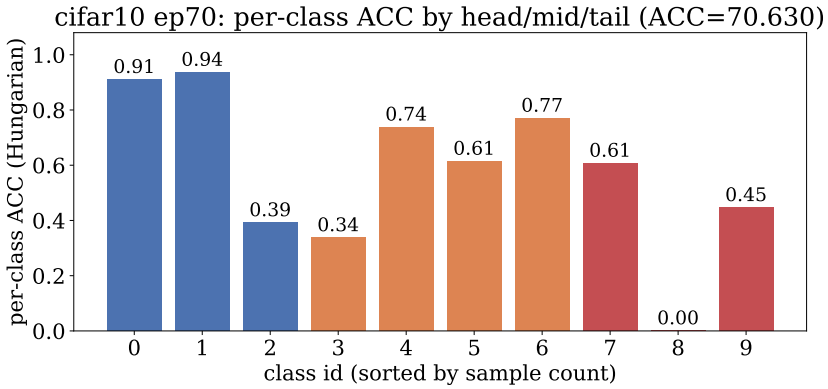


Figure 8: Per-class ACC on CIFAR-10 (imb_ratio = 10), at the reported epoch-110 checkpoint (Table 4). Bars show the per-class accuracy of every class (sorted by sample count), and the legend reports each group's mean per-class accuracy and sample count. In the figure, blue, orange, and red correspond to head, mid, and tail classes, respectively.

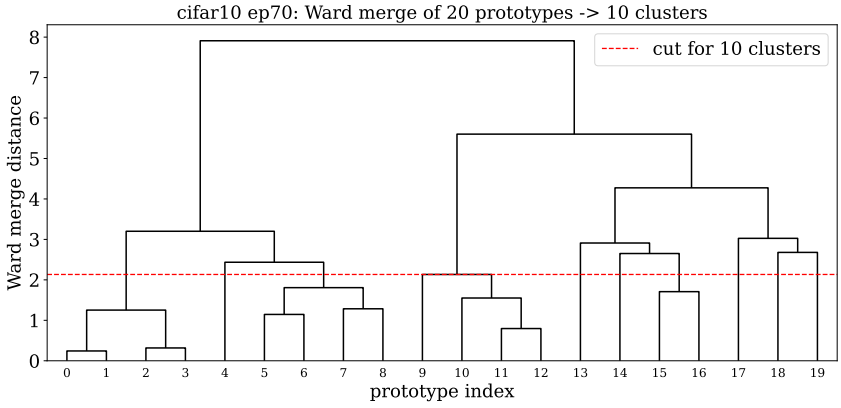


Figure 9: Ward merging of the $K = 20$ over-segmented prototypes into $K^* = 10$ clusters on CIFAR-10 (Eq. 12). The dashed red line marks the merge distance that yields exactly ten groups.

Figures 8 and 9 are drawn from the same epoch-110 checkpoint as Table 4. Figure 8 reports the per-class accuracy of all ten classes of CIFAR-10, sorted by sample count and colored by head/mid/tail groups; the title ACC is the class-count-weighted mean of the bars and CAA is their unweighted mean, so both are directly verifiable against Table 4 (70.6 and 67.5). Head and mid classes reach high per-class accuracy, while accuracy decays along the tail; the rarest class ($\approx$500 samples) remains the hardest case and is fully absorbed by a neighboring cluster, per-class ACC $= 0$, illustrating the residual tail difficulty that the long-tail-aware marginals (Sec. 3.1) and over-segmentation are designed to mitigate. The ACC–CAA gap in Table 4 reflects the same asymmetry: head classes dominate the class-count-weighted ACC, while CAA treats all classes equally.

Figure 9 illustrates the evaluation-time Ward merge of the $K = \rho K^* = 20$ over-segmented prototypes into $K^* = 10$ groups (Eq. 12). The dashed line marks the exact merge distance that yields ten groups, and prototypes merged below this distance are consolidated into one cluster.