GEOMETRY-PRESERVING STRUCTURED PRUNING FOR LONG-TAILED RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-tailed recognition models are trained on highly imbalanced data, where tail classes receive limited sample support. Structured channel pruning based on cross-sample importance may discard channels supporting tail-class regions, distorting local geometry and inter-class separation and thereby re-amplifying class bias during compression. To this end, we formulate long-tailed pruning as the preservation of class-local feature-space geometry and propose GB-DCS (Granular-Ball D-Optimal Channel Subspace Selection). GB-DCS uses class-specific granular balls to characterize local representation regions and temporarily masks candidate channels to measure signed deformations in their position, compactness, and separation from other classes. After balancing these responses across classes and frequency groups, GB-DCS applies a multi-kernel D-optimal criterion to retain channel subspaces with complementary geometric responses under a given structural budget. Experiments on moderate-scale CIFAR-100-LT and large-scale ImageNet-LT show improved overall recognition compared with existing methods, with substantial gains on tail classes. GB-DCS provides a geometry-aware framework for balancing model compactness and class-local representation preservation under long-tailed imbalance.

1 INTRODUCTION

Visual categories in real-world data occur at unequal frequencies: a small number of head classes provide abundant examples, whereas many tail classes have limited support (Liu et al., 2019). This imbalance affects both learned representations and classification (Cui et al., 2021; Kang et al., 2020). Under contrastive objectives, frequency imbalance can also distort class geometry (Yi et al., 2025; Zhu et al., 2022). Related analyses predict minority collapse under imbalanced training (Fang et al., 2021). However, model compression can further weaken representations of sparsely sampled classes, introducing an additional challenge to preserving their discriminative structure (Hooker et al., 2019).

Structured pruning removes channels and coupled parameters (Fang et al., 2023), but overall accuracy can conceal disproportionate damage to underrepresented examples (Hooker et al., 2019). Aggregate importance may overlook channels supporting small class regions. Data-free connectivity-based pruning (e.g., RReg) builds sparse networks without class-specific observations (Stewart et al., 2023), whereas tail-aware importance weighting (e.g., LTAP) adapts pruning priorities through class-performance feedback (Zhao et al., 2025). These criteria do not directly quantify class-local deformations or the geometric complementarity of retained channels. Figure 1 contrasts tail-aware importance ranking with geometry-based channel selection. We ask a representation-level question: *which channels should be retained to limit changes to class-local geometry under a structural budget?*

Channel removal can shift local centers, change dispersion, or alter inter-class separation. A scalar magnitude discards location and direction, while one prototype per class can conceal local variation. We therefore use granular balls (Xia et al., 2019) and signed masking responses to describe local effects and distinguish channels supporting similar or complementary geometric modes.

We introduce **Granular-Ball D-Optimal Channel Subspace Selection (GB-DCS)**. It measures removal effects on class-local regions, calibrates their relative contributions, and selects complementary channels across frequency groups (Figure 2).

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.

Our contributions are twofold:

- We formulate long-tailed structured pruning as preservation of **class-local geometry** under a structural budget, making changes in local position, compactness, and inter-class separation explicit.
- We propose a **geometry-aware selection criterion that accounts for class balance and channel complementarity**, providing a reusable basis for preserving class representations and reducing channel redundancy in long-tailed model compression.

2 Related Work

Long-tailed learning and compression. Balanced Softmax corrects class-frequency effects (Ren et al., 2020), LDAM uses class-dependent margins (Cao et al., 2019), and PLOT applies conflict-averse optimization (Zhou et al., 2024) to address training imbalance. Compression introduces an additional source of unequal errors across underrepresented examples (Hooker et al., 2019). LTAP adjusts pruning priorities to protect tail-class performance (Zhao et al., 2025), while theory connects pruning with feature separation in imbalanced contrastive learning (Liao et al., 2026). GB-DCS addresses class imbalance during channel selection, complementing the training objectives above.

Representation geometry and local regions. Minority collapse under imbalanced training (Fang et al., 2021) motivates attention to class separation. PaCo uses learnable class centers (Cui et al., 2021) and BCL balances class contributions (Zhu et al., 2022) in contrastive learning; FeatRecon reconstructs feature supports to improve separation (Yi et al., 2025). These works shape representations during learning. Granular-ball computing summarizes samples as labeled regions (Xia et al., 2019). GB-DCS uses class-local balls with fixed memberships to compare pruning effects across regions while retaining within-class variation.

Pruning criteria and feature reconstruction. Filter norms (Li et al., 2017) and loss-sensitive estimates (Molchanov et al., 2019) score structural importance. RReg instead builds sparse connectivity from random regular expander graphs at initialization, without training data (Stewart et al., 2023). Channel reconstruction preserves intermediate outputs (He et al., 2017), and SNOWS optimizes reconstruction through deeper nonlinear representations (Lucas & Mazumder, 2025). GB-DCS measures signed class-local deformations and calibrates class contributions, linking structural importance to class-aware geometric preservation. DepGraph supplies the dependency constraints for valid removal (Fang et al., 2023), within which GB-DCS selects channels.

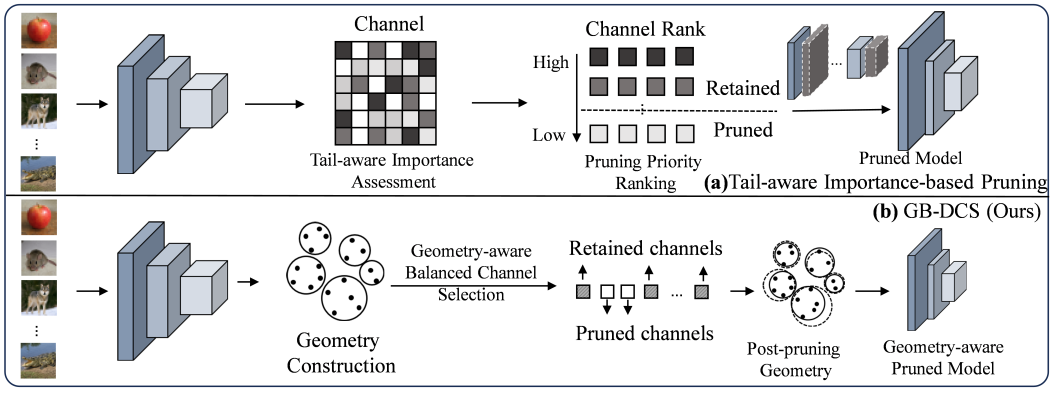

Figure 1: Tail-importance-aware and geometry-preserving pruning. **(a) Tail-importance-aware pruning (top):** channels are ranked by tail-aware importance to determine retention or removal. **(b) GB-DCS (bottom):** balanced geometric responses guide complementary channel selection for local geometry preservation.

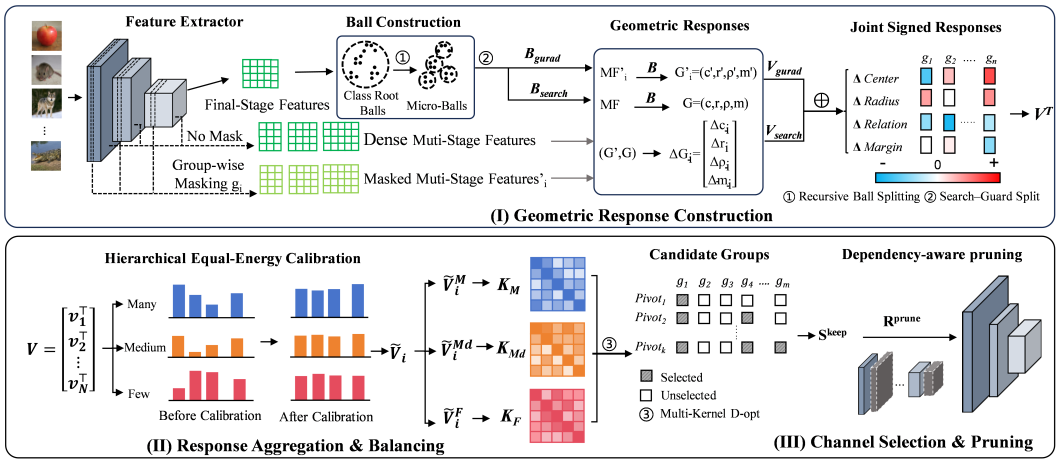

Figure 2: Overview of GB-DCS. **(I)** Class-wise balls define fixed local regions; dense and masked multi-stage features yield signed center, radius, relation, and margin changes per search/guard fold. **(II)** Fold responses are concatenated and calibrated across classes. **(III)** Three frequency-specific kernels guide complementary D-optimal selection and dependency-aware pruning. MF: multi-stage features; $\oplus$: concatenation; ΔG: scaled changes in Equation (5).

Complementary subset selection. SOSP accounts for correlations between structures when estimating pruning sensitivity (Nonnenmacher et al., 2022). Determinantal point processes combine item quality with subset diversity (Kulesza & Taskar, 2012), and Diversity Networks uses them for neuron selection and fusion (Mariet & Sra, 2016). GB-DCS defines candidate similarity through calibrated geometric responses in three frequency-specific kernels. Their mean and minimum scores evaluate set-level complementarity while attending to the least-covered frequency group.

3 Method

3.1 Problem Formulation

For a dense model f, we build class-local balls from normalized final-stage features of training probes and seek a dependency-valid removal set R that preserves their geometry:

$$\begin{aligned} R^\star &\in \arg\min_R \mathcal{D}_{\mathrm{bal}}(f_R, f), \\ \text{s.t.} \quad & |R \cap \mathcal{G}_h| = p_h \quad \forall h, \\ & \mathrm{FLOPs}(f_R) \le \beta\, \mathrm{FLOPs}(f). \end{aligned} \tag{1}$$

Here f_R masks R before export, and FLOPs refer to the exported structure. The groups $\{\mathcal{G}_h\}$ partition the candidate index set, with removal quotas p_h fixed by a preset budget profile; β is the retained-FLOPs budget. $\mathcal{D}_{\mathrm{bal}}$ measures class-balanced geometric change and is defined below. We perform multi-kernel D-optimal selection on balanced individual removal responses as a tractable surrogate objective for joint geometric preservation (Figure 2).

3.2 Geometric Response Construction

Input features and fixed memberships. For training probe i, let $\mathbf{h}_{iL}^0$ denote the spatially averaged output of the dense model's final feature stage L, before the classifier; 0 denotes the dense state. Its normalized vector $\mathbf{F}_i$ is the input to ball construction. Here $\mathrm{Normalize}$ denotes numerically stabilized L_2 normalization (Appendix B). Let I_c contain the probe indices of class c before assignment to the two folds $q \in \{\mathrm{search}, \mathrm{guard}\}$. We construct class-wise balls and then split them into folds:

$$\mathbf{F}_i = \mathrm{Normalize}(\mathbf{h}_{iL}^0), \qquad \mathcal{B} = \bigcup_c \mathrm{Partition}\big(\{(i, \mathbf{F}_i) : i \in I_c\}\big),$$
$$(\mathcal{B}_{\mathrm{search}}, \mathcal{B}_{\mathrm{guard}}) = \mathrm{SplitBalls}(\mathcal{B}). \tag{2}$$

Each ball $b \in \mathcal{B}_q$ has label y_b and members I_b that remain fixed across stages and masks. Search and guard are sample-disjoint folds used jointly for channel selection. Fold assignment, including classes with a single ball, is specified in Appendix B.

Each class is refined into local balls by recursively splitting eligible leaves under minimum-size and radius-contraction conditions, subject to depth and ball-count limits (Appendix B). The construction radius is the mean distance to the mean center; the response radius below is RMS.

A candidate g is a channel-removal unit coupled through the dependency graph. Its temporary mask is denoted $-g$. Write $\mathbf{h}_{i\ell}^t$ for the spatially averaged feature at stage ℓ under $t \in \{0, -g\}$.

Ball geometry from normalized features. For a fixed fold q and stage ℓ, suppress these indices until aggregation. Normalize each pooled sample before computing ball geometry:

$$\mathbf{z}_i^t = \mathrm{Normalize}(\mathbf{h}_i^t), \qquad \boldsymbol{\mu}_b^t = \frac{1}{|I_b|} \sum_{i \in I_b} \mathbf{z}_i^t,$$
$$(r_b^t)^2 = \frac{1}{|I_b|} \sum_{i \in I_b} \|\mathbf{z}_i^t - \boldsymbol{\mu}_b^t\|_2^2. \tag{3}$$

Normalized center directions $\bar{\boldsymbol{\mu}}_b^t = \mathrm{Normalize}(\boldsymbol{\mu}_b^t)$ are used for cosine relations and distances. Write $d_{ba}^t = \|\bar{\boldsymbol{\mu}}_b^t - \bar{\boldsymbol{\mu}}_a^t\|_2$. For fixed b, write $a^t := a_b^t$. The nearest different-class ball, relation, and geometric margin are

$$a^t = \arg\min_{a : y_a \neq y_b} d_{ba}^t,$$
$$\rho_b^t = \langle \bar{\boldsymbol{\mu}}_b^t, \bar{\boldsymbol{\mu}}_{a^0}^t \rangle, \tag{4}$$
$$m_b^t = d_{ba^t}^t - r_b^t - r_{a^t}^t.$$

The relation compares the fixed dense pair (b, a^0), while the margin tracks the current nearest competing region a^t. Center displacement and RMS radii use mean centers; neighbor distances use normalized center directions. Thus, the margin combines directional separation and local dispersion.

Signed changes under a candidate mask. Center and radius changes capture position and compactness; relations and margins capture inter-class separation. For a fixed ball, suppress b in the four response components:

$$\boldsymbol{\Delta}_g^c = \frac{\boldsymbol{\mu}^{-g} - \boldsymbol{\mu}^0}{d_0}, \qquad \Delta_g^r = \log \frac{r^{-g} + \epsilon_r}{r^0 + \epsilon_r},$$
$$\Delta_g^\rho = \rho^{-g} - \rho^0, \qquad \Delta_g^m = \mathrm{clip}_{[-10,10]}\left(\frac{m^{-g} - m^0}{|m^0| + \epsilon_m}\right). \tag{5}$$

The dense nearest-neighbor distance scale d_0 makes center displacement comparable across stages; ϵ_r, ϵ_m stabilize small radii and margins. These scales are fixed from each fold's dense geometry at each stage (Appendix B). Positive radius and relation responses indicate expansion and increased alignment with another class; negative margin responses indicate reduced separation. Center coordinates retain displacement direction.

3.3 Response Aggregation and Balancing

Aggregation across stages and folds. Restore ℓ, q here. For $k \in \{c, r, \rho, m\}$, let $\Delta^k_{g,\ell q}$ collect component k over all balls, with $n_{\ell qk}$ scalar coordinates. We aggregate the stage-wise blocks within each fold and concatenate the two folds into the combined geometric response $\mathbf{v}_g$:

$$\begin{aligned} \mathbf{v}_g^q &= \bigoplus_{\ell,k} \sqrt{\frac{\omega_\ell \gamma_k}{n_{\ell qk}}}\, \mathrm{vec}(\Delta^k_{g,\ell q}), \\ \mathbf{v}_g &= \frac{\mathbf{v}_g^{\mathrm{search}} \oplus \mathbf{v}_g^{\mathrm{guard}}}{\sqrt{2}}. \end{aligned} \tag{6}$$

The nonnegative weights ω_ℓ, γ_k control stage and component contributions. Coordinate normalization prevents larger blocks from dominating through dimension alone, and concatenation retains both folds without sign cancellation. Stacking candidate rows gives $\mathbf{V} = [\mathbf{v}_1^\top; \ldots; \mathbf{v}_N^\top]$, where $N = |\mathcal{G}|$ and $\mathcal{G} = \bigcup_h \mathcal{G}_h$.

Class-wise energy calibration. Let Π_c select coordinates of class c, and $s(c)$ denote its frequency group in $\mathcal{S} = \{\mathrm{many}, \mathrm{medium}, \mathrm{few}\}$. Normalize class energy over all candidates:

$$\begin{aligned} e_c^2 &= \frac{1}{N} \sum_{g \in \mathcal{G}} \|\Pi_c \mathbf{v}_g\|_2^2, \qquad a_c = \frac{1}{\max(e_c, \tau_c)}, \\ T_{jj} &= a_{c(j)}, \qquad\qquad \widetilde{\mathbf{v}}_g = T \mathbf{v}_g. \end{aligned} \tag{7}$$

Here T is diagonal, $c(j)$ identifies the coordinate's class, and $\widetilde{\mathbf{v}}_g$ is the calibrated response. Median-based floors τ_c limit amplification of nearly zero energy (Appendix B). The calibration factors are estimated over all candidates and fixed across dependency groups, preserving signs and relative candidate variation within each class. This defines $\mathcal{D}_{\mathrm{bal}}(f_R, f) = \|T\mathbf{v}_R\|_2^2$ in Equation (1): $\mathbf{v}_R$ measures the joint mask using the same procedure, with T fixed from individual responses. Class balancing is implemented by energy calibration; frequency-group balancing enters through the joint kernel objective below.

3.4 Channel Selection and Pruning

Three quality-weighted kernels. Fix dependency group h and suppress its index on all group-specific quantities below. Restrict the balanced candidate response $\widetilde{\mathbf{v}}_g$ to frequency group s:

$$E_{g,s} = \|\Pi_s \widetilde{\mathbf{v}}_g\|_2^2, \qquad \mathbf{u}_{g,s} = \mathrm{Normalize}(\Pi_s \widetilde{\mathbf{v}}_g). \tag{8}$$

Normalized directions encode response similarity. To retain response strength while discounting contraction and cross-fold inconsistency, assign the quality

$$w_{g,s} = \max\left\{ w_{\min}, \frac{\min\{\sqrt{E_{g,s}/M_s}, 5\}}{1 + \beta_N \widehat{N}_g + \beta_A A_g} \right\}. \tag{9}$$

M_s is the numerically stabilized positive median of energies for frequency group s within the current dependency group; $\widehat{N}_g$ is median-normalized radius contraction; A_g is the relative energy disagreement between folds. $\beta_N, \beta_A \geq 0$ control these penalties and $w_{\min} > 0$ prevents zero quality. Their complete definitions are in Appendix B. With U_s stacking rows $\mathbf{u}_{g,s}^\top$ and $W_s = \mathrm{diag}(w_{g,s} : g \in \mathcal{G}_h)$, use the quality–similarity factorization of determinantal kernels (Kulesza & Taskar, 2012) with a positive ridge:

$$K_s = W_s U_s U_s^\top W_s + \lambda I, \qquad s \in \mathcal{S}. \tag{10}$$

The ridge $\lambda > 0$ makes all principal submatrices positive definite.

Complementary selection under fixed quotas. A retained subset $S \subseteq \mathcal{G}_h$ determines the selected channel subspace through its retained coordinates. Define $L_s(S) = \log\det K_s[S, S]$ and $L_s(\varnothing) = 0$. The combined coverage objective is

$$J(S) = \frac{1-\alpha}{3} \sum_{s \in \mathcal{S}} L_s(S) + \alpha \min_{s \in \mathcal{S}} L_s(S). \tag{11}$$

The weight $\alpha \in [0,1]$ balances average coverage across frequency groups and coverage of the least-covered frequency group. The three kernels capture response complementarity within each frequency group, while the minimum term discourages neglecting any frequency group. This coverage objective is a surrogate for joint geometric preservation (Appendix C).

Starting from $S_0 = \varnothing$, greedily select $k_h = |\mathcal{G}_h| - p_h$ candidates and prune their complement:

$$\begin{aligned} g_t &\in \arg\max_{g \in \mathcal{G}_h \setminus S_t} J(S_t \cup \{g\}), \\ S_{t+1} &= S_t \cup \{g_t\}, \\ S_h^{\text{keep}} &= S_{k_h}, \\ R &= \bigcup_h (\mathcal{G}_h \setminus S_h^{\text{keep}}). \end{aligned} \tag{12}$$

The resulting R instantiates the removal set in Equation (1) and the joint response $\mathbf{v}_R$, with $|R \cap \mathcal{G}_h| = p_h$ by construction. Pivoted log-determinant updates (Chen et al., 2018) implement this greedy rule, with candidate indices breaking ties. The dependency graph maps the complement R (labeled R^{prune} in Figure 2) to coupled parameter slices, defining the pruning mask. Recovery training keeps these slices at zero. After recovery and any applicable classifier adaptation, physical removal of the masked channels and coupled parameters produces the structurally pruned network.

4 Experiments

We evaluate recognition accuracy and retained FLOPs on two long-tailed benchmarks, followed by paired ablations of local regions, response balancing, and channel selection.

4.1 Experimental Setup

Datasets. We evaluate moderate-scale CIFAR-100-LT (Cao et al., 2019; Krizhevsky, 2009) at $\text{IR} = 100, 50$ and large-scale ImageNet-LT (Liu et al., 2019), both with class-balanced test sets. Frequency groups use training counts: many-shot ($n_c > 100$), medium-shot ($20 \le n_c \le 100$), and few-shot ($n_c < 20$). Dataset statistics and sampling details appear in Appendix A.

Backbones and pruning protocol. We use ResNet-32/ResNet-50 (He et al., 2016) for CIFAR-100-LT/ImageNet-LT, with Balanced Softmax (BS) (Ren et al., 2020), LDAM-DRW (Cao et al., 2019), and PLOT-based training (Zhou et al., 2024) (DBLP in the tables). Recovery preserves the host objective and pruning mask; physical compression follows recovery and any classifier adaptation. The total probe budget across the search and guard folds is at most 1,200 training images for CIFAR-100-LT and 4,096 for ImageNet-LT. Training schedules and optimizer settings appear in Appendix A.

Baselines and metrics. We compare RReg (Stewart et al., 2023), LTAP (Zhao et al., 2025), and GB-DCS within each training method, using the corresponding dense model as a reference. Tables report overall top-1 accuracy (All), mean class accuracy within frequency groups (Head/Medium/Tail), retained FLOPs F, and accuracy retention C_{ref}:

$$F = 100\frac{\text{FLOPs}_{\text{pruned}}}{\text{FLOPs}_{\text{dense}}}, \qquad C_{\text{ref}} = 100\frac{A_{\text{pruned}}}{A_{\text{dense}}}.$$

Here A denotes All accuracy; all metrics are percentages. Gray rows are dense references; best accuracy values among pruned methods within each host are bold, including ties at one decimal place. GB-DCS reports single checkpoints and configurations selected by test performance; training and selection budgets differ across methods.

Classifier evaluation. Following decoupled evaluation (Kang et al., 2020; Sun et al., 2025), selected runs freeze encoder parameters after recovery and fine-tune the classifier with class-balanced sampling. The pruning structure remains fixed. GB-DCS hyperparameters, classifier-adaptation runs, and batch-normalization (BN) settings are listed in Appendix A.

Table 1: CIFAR-100-LT results with ResNet-32.

Method (IR = 100)	$F(\%)\downarrow$	Head↑	Medium↑	Tail↑	All↑	$C_{\text{ref}}(\%)\uparrow$
BS	100.0	62.6	48.5	27.0	47.2	100.0
+ RReg	35.1	53.0	36.3	19.0	37.3	79.0
+ LTAP	22.6	55.8	44.7	23.2	42.4	89.8
+ **GB-DCS**	22.6	**56.6**	**45.2**	**28.0**	**44.0**	**93.3**
LDAM-DRW	100.0	65.1	48.1	20.1	45.8	100.0
+ RReg	35.1	54.3	37.8	16.7	37.6	82.1
+ LTAP	24.8	**56.9**	**40.2**	18.8	39.8	86.9
+ **GB-DCS**	24.8	56.4	40.1	**20.9**	**40.0**	**87.4**
DBLP	100.0	61.4	46.9	23.6	45.3	100.0
+ RReg	35.1	52.1	39.2	17.5	37.5	82.8
+ LTAP	24.0	54.7	43.3	25.8	42.0	92.7
+ **GB-DCS**	24.0	**57.1**	**43.5**	**26.9**	**43.3**	**95.5**
Method (IR = 50)	$F(\%)\downarrow$	Head↑	Medium↑	Tail↑	All↑	$C_{\text{ref}}(\%)\uparrow$
BS	100.0	62.3	46.1	37.0	51.2	100.0
+ RReg	35.1	51.3	35.6	23.8	40.2	78.5
+ LTAP	22.6	**57.6**	43.4	34.1	47.8	93.4
+ **GB-DCS**	22.6	56.0	**44.7**	**36.9**	**47.9**	**93.6**
LDAM-DRW	100.0	64.5	43.0	26.4	49.1	100.0
+ RReg	35.1	52.9	39.8	23.7	42.4	86.4
+ LTAP	24.8	**58.8**	39.9	23.3	44.8	91.2
+ **GB-DCS**	24.8	56.5	**40.9**	**28.9**	**45.1**	**91.9**
DBLP	100.0	61.2	46.5	32.3	50.2	100.0
+ RReg	35.1	50.8	40.4	24.1	43.8	87.3
+ LTAP	24.0	**56.1**	43.5	31.5	46.7	93.0
+ **GB-DCS**	24.0	55.5	**44.4**	**35.3**	**47.4**	**94.3**

Table 2: ImageNet-LT results with ResNet-50.

Method	$F(\%)\downarrow$	Head↑	Medium↑	Tail↑	All↑	$C_{\text{ref}}(\%)\uparrow$
BS	100.0	60.9	48.8	32.1	51.0	100.0
+ RReg	52.1	41.1	36.0	18.2	35.5	69.6
+ LTAP	30.6	**58.5**	**45.0**	30.1	48.1	94.3
+ **GB-DCS**	30.6	**58.5**	**45.0**	**30.2**	**48.2**	**94.5**
LDAM-DRW	100.0	60.4	46.9	30.7	49.8	100.0
+ RReg	52.1	38.8	36.8	18.7	35.1	70.5
+ LTAP	30.8	57.8	41.9	23.3	45.4	91.2
+ **GB-DCS**	30.8	**59.5**	**45.4**	**31.4**	**48.9**	**98.2**
DBLP	100.0	61.7	47.1	30.3	50.4	100.0
+ RReg	52.1	40.0	35.7	18.8	35.0	69.4
+ LTAP	30.0	**58.5**	**45.2**	23.0	**47.3**	93.8
+ **GB-DCS**	30.0	**58.5**	**45.2**	**23.2**	**47.3**	**93.9**

4.2 Main Results

CIFAR-100-LT. Across all three training objectives at both imbalance ratios, GB-DCS has higher reported tail accuracy than LTAP at the same tabulated retained FLOPs (Table 1). With BS at IR=100, GB-DCS reaches 44.0% All and 28.0% Tail at 22.6% retained FLOPs, exceeding LTAP by 1.6 and 4.8 percentage points, respectively. At IR=50, GB-DCS improves both overall and tail accuracy, with gains of 0.1–0.7 and 2.8–5.6 percentage points, respectively.

ImageNet-LT. The gains vary across training objectives (Table 2). With LDAM-DRW, GB-DCS reaches 48.9% All and 31.4% Tail at 30.8% retained FLOPs, exceeding the reported LTAP results by 3.5 and 8.1 points. With classifier adaptation, BS and DBLP are close to LTAP: All differs by at most 0.1 points and Tail by at most 0.2 points.

4.3 Ablation Study

Number of granular balls. Across the tested $n_{\min}$ values, actual ball counts range from 111 to 381 while All accuracy varies by at most 0.33 points (Table 3); other sensitivity results appear in Appendix E.

Table 3: Ball-count sensitivity through minimum child size $n_{\min}$ under the paired protocol of Table 4. Counts precede fold splitting; best accuracies are bold.

$n_{\min}$	Actual balls	Head	Medium	Tail	All
2 (default)	381	**56.74**	45.20	28.60	**44.26**
3	240	55.29	44.66	**29.83**	43.93
4	199	56.29	45.23	28.60	44.11
6	111	**56.74**	**45.29**	28.40	44.23

Paired comparison. To examine component effects under matched conditions, the four variants on CIFAR-100-LT (BS, IR=100) share one epoch-100 dense checkpoint, remove 426/608 candidates (22.6% retained FLOPs), and recover to epoch 200 without classifier adaptation. This single-seed comparison evaluates the fixed final checkpoint on all 10,000 test images; further details and checkpoint-selected ablations appear in Appendix A.

Local-region modeling (Section 3.2): one prototype per class replaces multiple local regions to test the value of within-class structure.

Response balancing (Section 3.3): setting calibration weights to one tests how response-scale adjustment affects performance across frequency groups.

Complementary selection (Section 3.4): diagonal kernels retain individual weights but remove off-diagonal similarities to test candidate relationships.

Table 4: Paired ablations on CIFAR-100-LT (BS, IR=100; 22.6% retained FLOPs). Single seed, fixed epoch 200. Best values are bold.

Variant	Head	Medium	Tail	All
GB-DCS	56.74	**45.20**	**28.60**	**44.26**
One prototype per class	55.97	43.86	27.77	43.27
w/o response balancing	**57.03**	44.69	27.47	43.84
Diagonal kernels	55.46	44.89	27.40	43.34

Table 5: Selected layer3 diagnostics. Minimum center displacement and absolute radius change are bold.

	Before recovery	Epoch 200		
Variant	All: Center	All: Radius	Head: Margin	Tail: Margin
GB-DCS	0.8235	**0.0241**	-0.0112	$+0.0008$
One prototype per class	0.9151	0.0256	-0.0176	$+0.0051$
w/o response balancing	**0.8131**	0.0247	-0.0109	-0.0029
Diagonal kernels	0.8541	0.0249	-0.0078	-0.0002

Component effects. GB-DCS improves All/Tail accuracy by 0.99/0.83 points over one prototype per class, supporting local-region modeling (Table 4). Removing balancing lowers All/Tail by 0.42/1.13 points while increasing Head by 0.29 points. Compared with diagonal kernels, GB-DCS improves All/Tail by 0.92/1.20 points, supporting complementary selection.

Local-region preservation. Table 5 summarizes selected layer3 geometric diagnostics. Center and Radius denote displacement and absolute radius change; Margin is the change from the dense-model margin. Before recovery, GB-DCS yields smaller center displacement than the one-prototype and diagonal-kernel variants. After recovery, it yields the smallest average absolute radius change.

Balancing and tail separation. At the fixed final checkpoint, GB-DCS has greater tail-class center displacement than the variant without balancing, but its tail-class margin change is $+0.0008$ rather than -0.0029 (Table 11), and its tail accuracy is 1.13 percentage points higher (Table 4). These observations indicate that positional changes and separation margins should be considered together. Full measurements appear in Appendix D.

4.4 Discussion

Focus of pruning criteria. RReg and LTAP use sparse connectivity design and tail-aware importance, respectively, to reduce model cost while retaining recognition ability (Stewart et al., 2023; Zhao et al., 2025). Under long-tailed imbalance, class representations may differ in their sensitivity to pruning. GB-DCS therefore introduces class-local geometry as a complementary selection criterion to account for representation preservation across frequency groups.

Mechanism and observed effects. The paired ablations indicate that local regions, response balancing, and channel complementarity contribute to overall and tail accuracy. Together, these components connect class-local representation preservation with channel selection under a structural budget.

Approximation of joint pruning. Responses are measured one candidate at a time around the dense model. Joint removal can introduce nonlinear interactions or cancellation that the determinantal coverage objective does not predict (Appendix C). Fixed balls also describe only regions represented in the training probes, so poorly sampled class modes may be missed. A possible extension is to refresh selected responses and local regions using a small number of joint-mask probes; this would require evaluating both preservation accuracy and additional selection cost.

Budget allocation and computational cost. Fixed dependency-group quotas determine how many channels each group retains; GB-DCS optimizes their identities without reallocating capacity across groups. Moreover, candidate-wise masked evaluations add selection work as the candidate pool grows. Joint geometry-based budget allocation and coarse candidate screening are possible extensions, with the latter requiring care to retain channels supporting rare regions.

5 Conclusion

We presented GB-DCS, a structured channel pruning method that formulates long-tailed model compression as the preservation of class-local feature geometry. It combines granular-ball modeling, balanced geometric responses, and multi-kernel D-optimal selection to retain complementary channels under a structural budget. Experiments on CIFAR-100-LT and ImageNet-LT show overall and tail recognition gains at the reported FLOPs budgets. These results highlight class-local geometry and frequency balancing as useful criteria for reconciling model compression with tail recognition.

AI USE STATEMENT

Generative AI tools assisted with language polishing, translation, drafting and revising parts of the manuscript, and improving its organization. They also supported literature searches, organization of related work, and reference formatting. For methodological development and implementation, these tools helped draft and refine pseudocode, translate granular-ball design ideas into code, implement selected components from related methods, and check consistency between code, equations, and descriptions. Additional uses included feedback on methodology and experiments, inspection of code and archived logs, interpretation of recorded results, and LaTeX formatting. All reported experimental measurements come from model training and evaluation. The authors are responsible for research decisions, experimental execution, verification of AI-assisted code, mathematical expressions, references, and claims, and the final content of the submission.

REPRODUCIBILITY STATEMENT

Appendix A records the route-construction and training settings from the implementation and saved configurations. Appendices B–C supplement the construction, normalization, calibration, and determinant formulas. Section 4.1 records the evaluated configuration, model-selection procedure, and reporting conditions. Section 4.3 reports paired component ablations; Appendix E provides single-factor sensitivity results.

REFERENCES

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, pp. 1567–1578, 2019.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5622–5633, 2018.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 715–724, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. DepGraph: Towards any structural pruning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16091–16101, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *IEEE International Conference on Computer Vision*, pp. 1389–1397, 2017.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient ConvNets. In *International Conference on Learning Representations*, 2017.

Haixu Liao, Yating Zhou, Songyang Zhang, Meng Wang, and Shuai Zhang. Theoretical analysis of contrastive learning under imbalanced data: From training dynamics to a pruning solution. In *International Conference on Learning Representations*, volume 2026, pp. 138739–138789, 2026.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2537–2546, 2019.

Ryan Lucas and Rahul Mazumder. Preserving deep representations in one-shot pruning: A Hessian-free second-order optimization framework. In *International Conference on Learning Representations*, volume 2025, pp. 35617–35641, 2025.

Zelda Mariet and Suvrit Sra. Diversity networks. In *International Conference on Learning Representations*, 2016.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11264–11272, 2019.

Manuel Nonnenmacher, Thomas Pfeil, Ingo Steinwart, and David Reeb. SOSP: Efficiently capturing global correlations by second-order structured pruning. In *International Conference on Learning Representations*, 2022.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4175–4186, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

James Stewart, Umberto Michieli, and Mete Ozay. Data-free model pruning at initialization via expanders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4519–4524, 2023.

Siyu Sun, Han Lu, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, and Junchi Yan. Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance. In *International Conference on Learning Representations*, volume 2025, pp. 26681–26697, 2025.

Shuyin Xia, Yunsheng Liu, Xin Ding, Guoyin Wang, Hong Yu, and Yuoguo Luo. Granular ball computing classifiers for efficient, scalable and robust learning. *Information Sciences*, 483: 136–152, 2019.

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 *International Conference on Learning Representations*, volume 2025, pp. 69619–69651, 2025.

Zhe Zhao, Haibin Wen, Pengkun Wang, Shuang Wang, Zhenkun Wang, Qingfu Zhang, and Yang Wang. Balancing model efficiency and performance: Adaptive pruner for long-tailed data. In *42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77723–77739, 2025.

Zhipeng Zhou, Liu Liu, Peilin Zhao, and Wei Gong. Pareto deep long-tailed recognition: A conflict-averse solution. In *International Conference on Learning Representations*, volume 2024, pp. 54500–54526, 2024.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6908–6917, 2022.

A Implementation Details

Dataset construction. CIFAR-100 contains 100 classes of 32×32 images, with 500 training and 100 test examples per class (Krizhevsky, 2009). Following the exponential long-tail construction (Cao et al., 2019), classes are ordered by decreasing training frequency and subsampled to $n_c = \lfloor 500\,\mathrm{IR}^{-(c-1)/99} \rfloor$, $c = 1, \ldots, 100$. At IR=100 and 50, the smallest class contains 5 and 10 examples, respectively; the balanced test set retains all 10,000 images. ImageNet-LT subsamples ImageNet (Deng et al., 2009; Russakovsky et al., 2015) into 1,000 classes and 115,846 training images, with 5–1,280 per class ($\mathrm{IR} = 256$). Its provided validation and test lists contain 20,000 and 50,000 images, respectively, balanced across classes (Liu et al., 2019). All response probes come from the training list; frequency groups are determined before probe sampling. Table 6 summarizes the training and implementation settings.

Table 6: Training and implementation settings. GB-DCS numerical parameters are listed separately in Table 7.

Setting	Value
Backbones	ResNet-32 (CIFAR-100-LT); ResNet-50 (ImageNet-LT)
Training / pruning epoch	200 / 100 (CIFAR); 100 / 50 (ImageNet-LT)
Batch size	128 (CIFAR); 256 (ImageNet-LT)
Optimizer	SGD; initial learning rate 0.1; momentum 0.9; weight decay 2×10^{-4}
Learning-rate schedule	CIFAR: step decay at 160/180; ImageNet-LT: dense warmup/step recipe, then cosine recovery from 0.02
Probe budget	1200 (CIFAR); 4096 (ImageNet-LT)
Probe allocation	At most 16 examples per class; many/medium/few allocation 30%/30%/40%, followed by class round-robin allocation
Ball memberships	Dense final-layer features; fixed across stages and masks
Response space	Sample-wise L_2 normalization; mean centers and RMS radii; normalized center directions for neighbor distances
Micro-ball refinement	Balanced farthest-pivot splits; mean-distance construction radius
Response components	Center, radius, relation, margin; weights are stage–component energy coefficients, not fixed output-energy fractions
Physical pruning	Dependency-aware channel removal under fixed group quotas
Recovery	Host objective; fixed pruning route

Classifier-adaptation settings. Classifier-only adaptation with class-balanced sampling applies to CIFAR DBLP at IR=100, all CIFAR hosts at IR=50, and ImageNet-LT BS/DBLP; other rows use recovery checkpoints. CIFAR DBLP IR=100 updates BN statistics; CIFAR IR=50 and ImageNet-LT DBLP freeze them. The latter uses learning rate 2×10^{-4} and many/medium/few loss weights $1.00/1.06/0.50$ (normalized across classes), reporting adaptation epoch 3. Representation comparisons require matched classifier and BN protocols.

Evaluated response convention. Tabulated runs L_2-normalize pooled samples before computing centroids and RMS radii, then additionally normalize centroids for neighbor distances and margins, as defined in Section 3.2.

Table 7: GB-DCS settings. The response-component order is center/relation/radius/margin; ball limits apply before the search/guard split.

Setting	CIFAR-100-LT	ImageNet-LT
$n_{\min}$ / ξ / maximum depth	2 / 0.985 / 3	
Maximum balls per class	4	
Global ball cap	400	4,000
Response-component weights	1 / 0.75 / 0.35 / 1.5	
Feature stages	layer1–3	layer2–4
Stage weights	0.50 / 0.35 / 0.15	
Center-response projection	None	32 dimensions/ball
(λ, α)	$(0.01, 0.65)$	$(10^{-6}, 0.25)$
(β_N, β_A)	$(0.35, 0.25)$	$(0.35, 0)$
$(w_{\min}, \eta)$	$(0.05, 0.05)$	

CIFAR BS/LDAM at IR=50 use $\beta_N = 0$, and LDAM at IR=50 uses $\lambda = 10^{-6}$. ImageNet also collects layer1 with zero weight.

Fixed-epoch paired ablations. All four variants start from the same epoch-100 dense model, prune 426/608 candidates, retain 22.59% FLOPs and 101,512 parameters, and recover to epoch 200 without classifier adaptation. Each final checkpoint is reloaded and evaluated on all 10,000 test images.

Supplementary checkpoint-selected ablations. Table 8 reports a supplementary paired comparison. Variants share the dense checkpoint, optimizer, training subset, random initial state, BS objective, 1,019 probe images, and adaptive-augmentation rule; augmentation subsequently follows each model's predictions. They prune 426/608 candidates under 17 dependency-group quotas and recover through epochs 101–200. Removing balancing sets all response calibration weights to one. Energy ranking retains the largest balanced response energies within each dependency group and omits both determinant selection and its quality penalties. For each row, all four metrics are evaluated at the checkpoint selected to maximize overall test accuracy. The reference initialization was chosen from preliminary runs. Table 9 gives paired results from another initialization, with a different ranking in overall accuracy. An additional unpaired full-method run achieved 43.2% All. Results are reported separately for each run.

Table 8: Supplementary checkpoint-selected component ablations on CIFAR-100-LT, BS, IR=100, at 22.6% retained FLOPs. Each row uses the same best-All checkpoint for all four metrics. Best values, including rounded ties, are bold.

Variant	Head	Medium	Tail	All
GB-DCS	56.6	**45.2**	**28.0**	**44.0**
w/o response balancing	57.3	44.5	27.6	43.9
Response energy Top-k	**57.5**	44.6	26.8	43.8

Table 9: Additional initialization on CIFAR-100-LT, BS, IR=100, at 22.6% retained FLOPs. Reporting follows Table 8.

Variant	Head	Medium	Tail	All
GB-DCS	55.4	46.1	**25.0**	43.0
w/o response balancing	56.6	46.1	23.7	43.1
Response energy Top-k	**56.9**	**46.5**	24.7	**43.6**

B Construction and Numerical Details

Numerical convention. We use $\varepsilon = 10^{-12}$ and write $\mathrm{Normalize}(\mathbf{x}) = \mathbf{x}/\max\{\|\mathbf{x}\|_2, \varepsilon\}$, which maps zero vectors to zero.

Class-wise ball construction. We spatially average each stage and construct memberships from normalized dense final-layer features, as in Equation (2). Each observed class starts with one root at depth zero. Sample order and probe sampling are fixed during construction; the depth and ball-count limits are listed in Appendix A.

For a class leaf set I, let $\mathbf{c}_I = |I|^{-1} \sum_{i \in I} \mathbf{F}_i$. Choose the point farthest from this mean and then the point farthest from that pivot; stably sort the distance differences to form balanced children:

$$\begin{aligned} u &= \arg\max_{i \in I} \|\mathbf{F}_i - \mathbf{c}_I\|_2, \qquad v = \arg\max_{i \in I} \|\mathbf{F}_i - \mathbf{F}_u\|_2, \\ \sigma &= \mathrm{argsort}_{\mathrm{stable}}\big(\|\mathbf{F}_i - \mathbf{F}_u\|_2 - \|\mathbf{F}_i - \mathbf{F}_v\|_2 : i \in I\big), \\ I_- &= \{\sigma_1, \ldots, \sigma_{\lfloor |I|/2 \rfloor}\}, \qquad I_+ = I \setminus I_-. \end{aligned} \tag{B.1}$$

The construction radius is the mean distance to the mean center, $R(I) = |I|^{-1} \sum_{i \in I} \|\mathbf{F}_i - \mathbf{c}_I\|_2$. For a leaf below the depth limit with $u \neq v$ and $R(I) > \varepsilon$, accept a split only when both children contain at least $n_{\min}$ samples and

$$\frac{|I_-| R(I_-) + |I_+| R(I_+)}{|I| R(I)} \leq \xi, \qquad \min\{|I_-|, |I_+|\} \geq n_{\min}. \tag{B.2}$$

A global priority queue orders eligible leaves using class frequency, parent radius, sample count, and radius contraction. Refinement stops at the depth, per-class, or global ball limit, or when no eligible leaf remains. Memberships stay fixed across stages and masks; the construction radius differs from the RMS response radius in Equation (3).

Unsplit leaves remain valid balls. After construction, balls sorted by identifier and member indices alternate between search and guard within each class. A sole ball with at least two members is split at $\lfloor n/2 \rfloor$ in member order; a singleton enters search only. Thus, the folds share no samples. Each geometry fold must represent at least two classes to define different-class neighbors.

Dense reference scales. As in Section 3.2, fix a fold and stage and omit their indices. Restoring the ball index, let $d_b^0 = d_{b a_b^0}^0$. The scales in Equation (5) are

$$\begin{aligned} d_0 &= \max\{\mathrm{median}_b\, d_b^0, 10^{-4}\}, \\ \epsilon_m &= 0.1 d_0 + 10^{-6}, \\ \epsilon_r &= 0.1 \max\{\mathrm{median}_b\, r_b^0, 10^{-4}\} + 10^{-6}. \end{aligned} \tag{B.3}$$

The reference distance uses normalized center directions, as do neighbor selection and cosine relations. RMS radii use the mean centers of normalized samples. With fixed memberships and $c_i > 0$, normalizing $c_i \mathbf{h}_i$ preserves the samples whenever both $\|\mathbf{h}_i\|_2$ and $c_i \|\mathbf{h}_i\|_2$ exceed the normalization floor. Applying this condition to both dense and masked states preserves the responses. This invariance does not apply to rebuilding memberships from rescaled raw features.

Optional center projection. For input feature dimension d and configured projection dimension $d' < d$, map each center-response row to $\mathbf{\Delta}_g^c P$, leaving the other response components unchanged. Here $P \in \{-1/\sqrt{d}, +1/\sqrt{d}\}^{d \times d'}$ is sampled once per stage and shared by all candidates in that stage. The normalizing count $n_{\ell q c}$ uses the projected dimension. Without projection it equals $|\mathcal{B}_q| d$; the other three blocks each contain $|\mathcal{B}_q|$ scalar coordinates.

Energy floors. All medians use the lower middle value for even counts, matching the code. med_+ additionally excludes nonpositive values and returns 1 if none remain. With floor ratio η, Equation (7) uses

$$\tau_c = \max\{\eta\, \mathrm{med}_+(e_{c'} : s(c') = s(c)), \varepsilon\}. \tag{B.4}$$

Absent floor activation, class scaling gives $a_c e_c = 1$, an aggregate identity over candidates rather than an equal-energy constraint on each candidate's response blocks. When the absolute floors in Equations (8)–(9) are inactive, a common positive scaling within a frequency group leaves its normalized directions and median-normalized energies unchanged.

Candidate quality statistics. Fix dependency group h as in Section 3.4. The scale M_s is the positive energy median within this group, floored at ε, as in Equation (9). Let $[x]_+ = \max(x, 0)$, applied elementwise. The radius-response block $\Delta^r_{g,\ell q}$ collects the radius responses in Equation (5) over balls, as in Section 3.3. Taking the mean over balls, we define the contraction statistic and its normalized value as

$$\begin{aligned} N_g &= \frac{1}{2} \sum_{q,\ell} \omega_\ell \operatorname{mean}\big([-\Delta^r_{g,\ell q}]_+\big), \\ \widehat{N}_g &= \operatorname{clip}_{[0,10]}\left(\frac{N_g}{\max\{\operatorname{med}_+(N_i : i \in \mathcal{G}_h), \varepsilon\}} \right). \end{aligned} \tag{B.5}$$

The relative cross-fold energy disagreement is

$$A_g = \frac{\big| \|\mathbf{v}_g^{\text{search}}\|_2^2 - \|\mathbf{v}_g^{\text{guard}}\|_2^2 \big|}{\|\mathbf{v}_g^{\text{search}}\|_2^2 + \|\mathbf{v}_g^{\text{guard}}\|_2^2 + \varepsilon}. \tag{B.6}$$

These quantities are contraction and consistency statistics, not measured accuracy damage or estimates of statistical noise variance. Quality weights use frequency-group-specific energies, and channel selection follows the three-kernel objective in Equation (11).

Greedy determinant updates. Within the fixed group h, set $d_{g,s}(\varnothing) = K_s[g, g]$. For a nonempty retained set S and an unselected candidate g, the Schur complement gives

$$\begin{aligned} d_{g,s}(S) &= K_s[g, g] - K_s[g, S] K_s[S, S]^{-1} K_s[S, g], \\ L_s(S \cup \{g\}) &= L_s(S) + \log d_{g,s}(S). \end{aligned} \tag{B.7}$$

The implementation maintains these residuals through pivoted updates, with machine-epsilon flooring for numerical stability. Candidate scoring evaluates the mean and minimum of the *updated cumulative* log determinants. It does not take the minimum of isolated increments for individual frequency groups. Selection stops at the exact retention quota, and unselected candidates form the removal route.

C Response Complementarity and Geometric Preservation

The kernel criterion asks whether retained channels cover distinct, sufficiently strong patterns of local change. For two nonzero unit response directions in one dependency group and frequency group, let $\cos\theta = \mathbf{u}_i^\top \mathbf{u}_j$. Their kernel determinant is

$$\det K[\{i, j\}, \{i, j\}] = (w_i^2 + \lambda)(w_j^2 + \lambda) - w_i^2 w_j^2 \cos^2\theta. \tag{C.1}$$

With quality weights fixed, the determinant increases as directions become less collinear. A retained set can therefore support a broader set of local geometric effects than a set chosen only for large individual response norms. Class calibration and kernel aggregation across frequency groups give weight to tail-related coordinates even when their unbalanced energy is weak.

Under the approximation $\mathbf{v}_R \approx \sum_{g \in R} \mathbf{v}_g$, joint damage would involve

$$\left\| \sum_{g \in R} \widetilde{\mathbf{v}}_g \right\|_2^2 = \sum_{g \in R} \|\widetilde{\mathbf{v}}_g\|_2^2 + 2 \sum_{\substack{i,j \in R \\ i<j}} \langle \widetilde{\mathbf{v}}_i, \widetilde{\mathbf{v}}_j \rangle. \tag{C.2}$$

Opposite vectors can cancel in this sum but are as collinear as equal directions in Equation (C.1). Section 4.3 and Appendix D report joint-mask geometry before recovery and geometry at the fixed final checkpoint alongside the paired recognition results.

D Geometry Diagnostics

Tables 10 and 11 report geometry before recovery and at epoch 200 for the paired runs in Table 4. All variants share dense-model reference memberships. Ball metrics are averaged within each class, then equally over classes in each partition. Head/Tail correspond to Many/Few in the recorded diagnostics. Center and Radius denote displacement and absolute radius change; Δm and Δm_{coord} are signed margin and same-coordinate margin changes (evaluated model minus dense model).

Table 10: Layer3 changes under the complete joint mask, before any recovery update.

Group	Variant	Center	Radius	Δm	$\Delta m_{\mathrm{coord}}$
All	GB-DCS	0.8235	0.2989	$+0.3349$	$+0.3734$
	One prototype per class	0.9151	0.3388	$+0.3999$	$+0.4383$
	w/o response balancing	0.8131	0.2757	$+0.2992$	$+0.3365$
	Diagonal kernels	0.8541	0.2957	$+0.3304$	$+0.3685$
Head	GB-DCS	0.8455	0.2897	$+0.3035$	$+0.3405$
	One prototype per class	0.9363	0.3318	$+0.3729$	$+0.4106$
	w/o response balancing	0.8380	0.2682	$+0.2689$	$+0.3052$
	Diagonal kernels	0.8704	0.2863	$+0.2977$	$+0.3343$
Medium	GB-DCS	0.8089	0.2979	$+0.3413$	$+0.3820$
	One prototype per class	0.9013	0.3361	$+0.4019$	$+0.4416$
	w/o response balancing	0.7990	0.2756	$+0.3085$	$+0.3471$
	Diagonal kernels	0.8430	0.2949	$+0.3380$	$+0.3779$
Tail	GB-DCS	0.8150	0.3108	$+0.3641$	$+0.4016$
	One prototype per class	0.9065	0.3501	$+0.4291$	$+0.4669$
	w/o response balancing	0.8005	0.2847	$+0.3238$	$+0.3605$
	Diagonal kernels	0.8481	0.3075	$+0.3598$	$+0.3975$

Table 11: Layer3 changes at the fixed epoch-200 checkpoint.

Group	Variant	Center	Radius	Δm	$\Delta m_{\mathrm{coord}}$
All	GB-DCS	0.3038	0.0241	-0.0030	$+0.0021$
	One prototype per class	0.2998	0.0256	-0.0041	-0.0005
	w/o response balancing	0.2990	0.0247	-0.0032	$+0.0004$
	Diagonal kernels	0.3011	0.0249	-0.0054	-0.0026
Head	GB-DCS	0.3026	0.0239	-0.0112	-0.0063
	One prototype per class	0.2965	0.0249	-0.0176	-0.0126
	w/o response balancing	0.2992	0.0244	-0.0109	-0.0123
	Diagonal kernels	0.2991	0.0254	-0.0078	-0.0097
Medium	GB-DCS	0.3117	0.0253	$+0.0019$	$+0.0096$
	One prototype per class	0.3089	0.0267	$+0.0014$	$+0.0082$
	w/o response balancing	0.3098	0.0270	$+0.0044$	$+0.0122$
	Diagonal kernels	0.3113	0.0259	-0.0075	$+0.0014$
Tail	GB-DCS	0.2960	0.0229	$+0.0008$	$+0.0032$
	One prototype per class	0.2931	0.0253	$+0.0051$	$+0.0035$
	w/o response balancing	0.2862	0.0222	-0.0029	$+0.0014$
	Diagonal kernels	0.2916	0.0232	-0.0002	$+0.0011$

E SINGLE-FACTOR HYPERPARAMETER SENSITIVITY

Protocol and reporting scope. The additional experiments use CIFAR-100-LT, BS, IR=100, and ResNet-32. Each changes one pruning-stage parameter from the reference in Table 7, using the same epoch-100 dense checkpoint, optimizer state, sampling state, and 1,019 training probes. All runs remove 426/608 candidates under the same 17 dependency-group quotas, retain 22.5918% FLOPs and 101,512 parameters, and recover through epochs 101–200 without classifier adaptation. The final checkpoint is reloaded and evaluated on all 10,000 test images; no best-epoch metric is substituted. We report 21 additional runs and the historical reference.

Construction granularity. Table 12 reports construction parameters, including the full frequency-group results for Table 3. Increasing $n_{\min}$ from 2 to 3/4/6 reduces actual ball counts from 381 to 240/199/111 and gives 43.93/44.11/44.23% All, respectively. The overall changes are small and non-monotonic; $n_{\min} = 3$ has the highest Tail accuracy in this comparison. All settings retain the per-class cap of four and maximum depth of three. These counts precede fold splitting: a class with one ball can be split by samples across search and guard. Lowering the global cap from 400 to 100/200/300 gives 43.76/43.50/43.41% All, compared with 44.26% for the reference; more balls do not imply monotonically higher accuracy.

Table 12: Single-factor construction sensitivity. Balls and median samples per ball are measured before fold splitting; swaps count newly removed candidates relative to the reference.

Setting	Balls	Median	Swaps	Head	Medium	Tail	All
Default	381	3.0	0	56.74	45.20	28.60	44.26
$n_{\min} = 3$	240	5.0	50	55.29	44.66	29.83	43.93
$n_{\min} = 4$	199	5.0	54	56.29	45.23	28.60	44.11
$n_{\min} = 6$	111	10.0	50	56.74	45.29	28.40	44.23
Cap $= 100$	100	10.0	50	56.14	44.51	28.43	43.76
Cap $= 200$	200	3.0	55	56.14	44.83	27.20	43.50
Cap $= 300$	300	3.0	47	56.11	44.09	27.80	43.41
$\xi = 0.90$	283	3.0	46	56.77	44.51	28.20	43.91
$\xi = 0.95$	381	3.0	0	56.74	45.20	28.60	44.26
$\xi = 1.00$	381	3.0	0	56.74	45.20	28.60	44.26

Split threshold. At $\xi = 0.95, 0.985, 1.00$, ball memberships, responses, and pruning routes are identical, with 44.26% All after recovery in each setting. The stricter $\xi = 0.90$ gives 283 balls and 43.91% All.

Stage weighting. In Table 13, equal and reversed stage weights reduce All by 0.42 and 0.14 points relative to the reference $(0.50, 0.35, 0.15)$. Layer1-only, layer2-only, and layer3-only responses reduce All by 0.92/0.37/0.99 points. The default multi-stage setting achieves the highest All accuracy in this comparison. Single-stage variants use the same group quotas, and some candidates may have zero response at the observed stage.

Table 13: Single-factor response and selection sensitivity. All settings construct 381 balls; stage weights follow layer1/layer2/layer3. Swaps count newly removed candidates.

Setting	Swaps	Head	Medium	Tail	All
Default	0	56.74	45.20	28.60	44.26
$\omega = (1/3, 1/3, 1/3)$	23	55.20	45.34	28.83	43.84
$\omega = (0.15, 0.35, 0.50)$	26	56.86	45.11	28.10	44.12
$\omega = (1, 0, 0)$	86	56.31	44.86	26.43	43.34
$\omega = (0, 1, 0)$	73	56.40	45.34	27.60	43.89
$\omega = (0, 0, 1)$	40	56.46	43.74	27.33	43.27
$\alpha = 0$	21	56.34	45.31	27.83	43.93
$\alpha = 0.25$	15	55.97	45.63	27.30	43.75
$\alpha = 1.00$	12	56.43	44.89	27.20	43.62
$\lambda = 0.1$	5	56.57	43.37	27.57	43.25
$\beta_N = 0$	39	55.71	44.51	28.63	43.67
$\beta_N = 0.175$	27	56.29	44.91	28.03	43.83
$\beta_N = 0.70$	20	54.71	45.03	27.20	43.07

Selection parameters. The tested $\alpha = 0, 0.25, 0.65, 1$ give 43.93/43.75/44.26/43.62% All. This parameter balances the mean and weakest cumulative kernel scores in Equation (11); it does not assign a fixed weight to the Tail group. We evaluate the ridge at $\lambda \in \{0.01, 0.1\}$. For the contraction penalty $\beta_N = 0, 0.175, 0.35, 0.70$, All is 43.67/43.83/44.26/43.07%. Here β_N weights the radius-contraction statistic in Equation (9), not a statistical noise estimate. Among the reported settings, the reference achieves the highest All accuracy. Increasing λ from 0.01 to 0.1 or β_N from 0.35 to 0.70 reduces All accuracy by 1.01 and 1.19 percentage points, respectively.

Complementary geometry diagnostics. Table 14 compares layer3 geometry and recognition results for the reference and $\beta_N = 0.70$. Both use the same dense-model reference balls and 5,000 test images. Accuracy on the full 10,000-image test set is reported in Table 13.

Table 14: Selected layer3 diagnostics, averaged equally over classes. Center is displacement; radius and same-coordinate Euclidean margin are absolute changes from the dense model. Accuracy uses the 5,000-image diagnostic subset.

Stage	Setting	Center	Radius	Margin	Acc. (%)
Before recovery	Default	0.8235	0.2989	0.3746	1.76
Before recovery	$\beta_N = 0.70$	0.8295	0.3398	0.4413	1.04
Epoch 200	Default	0.3038	0.0241	0.0562	44.38
Epoch 200	$\beta_N = 0.70$	0.3051	0.0237	0.0543	43.36