WHERE DOES OVERFITTING LIVE? DEPTH-DEPENDENT STRUCTURE IN PRUNED NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although structured pruning is known to improve generalization in over-parameterized networks, which parameters carry the overfitting burden—and why pruning consistently finds them—remains unclear. We show that, during training, the deepest active block undergoes a simultaneous collapse in both BN scale factors and channel L_2-norms that is absent in shallower blocks: overfitting-related capacity concentrates toward the network's deepest stage, which we term the *noise sink*. On ResNet-18, all three criteria exclusively target the deepest stage until the architecturally determined 53.33% threshold; on ResNet-34, exclusivity is criterion-dependent; on VGG-16, the two deepest stages are co-targeted from the first cuts. Crossing the stage-exhaustion boundary triggers a sharp, criterion-dependent generalization transition. Sequential block removal further reveals that the noise sink role is not fixed to any specific block: after the deepest block is removed, the next deepest undergoes the same dual collapse and becomes the new pruning target, confirmed on ResNet-18, ResNet-34, and VGG-16. Controlled experiments show that Block 4 suppression strengthens monotonically with training data volume, the L_2 component persists under SGD+momentum, and the pattern is invariant to label noise—indicating a general property of over-parameterized network convergence rather than a small-data or optimizer artifact. Code is available at `https://anonymous.4open.science/r/depth-structure-pruning-CB93/`.

1 INTRODUCTION

It is well established that over-parameterized models are prone to overfitting whenever model capacity substantially exceeds the effective complexity of the training task (Zhang et al., 2017). This overparameterized regime arises naturally when data is scarce, but also when the architecture is sufficiently large relative to the task. However, most existing works focus predominantly on macroscopic generalization errors or loss curves (Neyshabur et al., 2017), leaving the microscopic structural level of the network—such as how model parameters evolve and how internal basic blocks divide their functional labor during overfitting—largely underexplored.

Neural network pruning (Han et al., 2015; LeCun et al., 1989) is a method that compresses over-parameterized networks by eliminating redundant parameters with little to no degradation in model performance. Furthermore, research has demonstrated that under certain conditions, pruning can act as a regularization technique to enhance generalization (Liu et al., 2019). However, a vast majority of prominent pruning paradigms, including global magnitude pruning and the seminal Iterative Magnitude Pruning (IMP) used in the Lottery Ticket Hypothesis, often operate under the implicit assumption that network layers are homogeneous entities (Han et al., 2015; Frankle & Carbin, 2019). This approach obscures the profound functional differences between varying depths in feature extraction and noise fitting, preventing us from discerning the physical essence of why pruning improves generalization.

To bridge this gap, this paper examines two questions: (1) Where does overfitting capacity concentrate within the architecture? (2) What governs the critical transition from beneficial pruning to harmful structural damage? To answer these, we designed pruning experiments across various

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.

pruning ratios in the overparameterized regime, using a 10% CIFAR-10 subset as a convenient experimental setting that ensures strong overparameterization with standard architectures. We utilized three distinct pruning criteria (γ, L_2, and $\gamma \times L_2$) to compare the distribution of pruned channels across each layer. Additionally, we conducted block removal experiments by alternately training and removing the entire deepest block. We find that in the overparameterized regime, overfitting capacity concentrates primarily in the deepest block. Moreover, there exists a threshold determined by the model's topological architecture, which is associated with a sharp transition in model generalization.

Our main contributions are as follows:

- **Depth-dependent structural concentration.** In the overparameterized regime, the deepest active block undergoes a dual collapse in BN γ and channel L_2-norm, concentrating the network's residual memorization capacity. Block removal experiments confirm this concentrated stage is not merely redundant but carries active overfitting-related capacity; controlled label-noise experiments show the concentration is independent of label-noise level, and a randomized pruning control confirms preferential Block 4 targeting reflects learned weight structure, not the pruning procedure.
- **An architecturally determined, criterion-dependent threshold.** We identify a pruning ratio threshold tied to the channel fraction of the deepest block (53.33% for ResNet-18; 40.68% for ResNet-34), beyond which generalization undergoes a sharp transition. The sharpness of this transition depends on the pruning criterion: γ produces a gradual crossing, while magnitude-sensitive criteria produce a sharp event triggered by a single noise-sensitive channel in Block 2.
- **Noise sink migration and a structural perspective on IMP.** After the deepest block is removed, the same dual collapse emerges in the new deepest active block, showing the noise sink role is dictated by topological depth. IMP's known advantage over one-shot pruning is consistent with this migration: iterative retraining allows the network time to complete the upward transfer of memorization capacity before each pruning step.

Related Work. Due to space constraints, we defer a comprehensive discussion of related works—including **Structural Pruning** (Li et al., 2017; Liu et al., 2017), **Generalization in Compressed Networks** (Zhang et al., 2017; Bartoldson et al., 2020), and the **Lottery Ticket Hypothesis (LTH)** (Frankle & Carbin, 2019)—to Appendix A.

2 Experimental Setup

[1]

2.1 Dataset, Preprocessing and Hyperparameters

We use CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009). To create an overparameterized regime, we randomly sample 10% of each training set: for CIFAR-10, 4,000 training / 1,000 validation / 1,000 test images; for CIFAR-100, 40 images per class (4,000 training), with 1,000 validation and 1,000 test. Four random seeds (42–45) confirm robustness. Standard augmentation is applied: resize to 36×36, random crop to 32×32, random horizontal flip, color jitter (factor 0.2), and channel-wise normalization ($\mu = [0.4914, 0.4822, 0.4465]$, $\sigma = [0.2470, 0.2435, 0.2616]$).

Training uses Cross-Entropy loss and Adam (β_1=0.9, β_2=0.999), batch size 256, gradient clip 1.0. Pre-training: 100 epochs, lr= 1×10^{-4}, wd= 1×10^{-4}. Fine-tuning: 40 epochs, lr= 1×10^{-5}, wd= 1×10^{-5}. Pruning ratios: 20%–90% (12 levels), with 2% resolution between 50% and 60%.

2.2 Model Architecture and Pruning Strategy

We select ResNet-18 and ResNet-34 (He et al., 2016) to serve as our baseline model. For the pruning strategy, we exclusively prune the internal bottleneck channels within each BasicBlock: we remove the output channels of `conv1`, the corresponding parameters of `bn1`, and the input

[1] Code: `https://anonymous.4open.science/r/depth-structure-pruning-CB93/`

channels of `conv2`, while leaving the final output channels of `conv2`, the corresponding `bn2`, and the shortcut connection completely intact. This method effectively trims network parameters without disrupting the tensor dimension alignment between x and output $\mathcal{F}(x)$, thereby preserving the backbone topology of the shortcut connection.

We designed three importance evaluation criteria to score each channel in the network and then perform the global ranking on all importance scores. A specified proportion of channels with the lowest scores are our target channels to prune. Let $\mathbf{W}_j \in \mathbb{R}^{C_{\text{in}} \times K \times K}$ denote the weight tensor of the j-th output channel in a convolutional layer, where C_{in} and K represent the number of input channels and kernel size, respectively. Let γ_j be the corresponding scaling parameter in the subsequent Batch Normalization (BN) layer.

Pure L_2-Norm Criterion. According to the magnitude-based pruning assumptions (Li et al., 2017), this metric evaluates the static energy of the filter, considering channels with smaller weight norms as less informative:

$$S_{L_2}(j) = \|\mathbf{W}_j\|_2 = \sqrt{\sum_{c=1}^{C_{\text{in}}} \sum_{m=1}^{K} \sum_{n=1}^{K} (\mathbf{W}_{j,c,m,n})^2} \tag{1}$$

Pure Gamma (γ) Criterion. Treating the BN layer as a channel selector (Liu et al., 2017), this metric uses the scaling factor γ to measure the dynamic activation response. A near-zero $|\gamma_j|$ indicates that the channel was spontaneously suppressed during training:

$$S_{\gamma}(j) = |\gamma_j| \tag{2}$$

Joint $L_2 \times$ BN Criterion. To avoid misjudging channels where weight magnitude and BN scaling diverge (e.g., a large-weight channel suppressed by a tiny γ), we combine the L_2-Norm with the BN scaling factor γ:

$$S_{\text{Product}}(j) = \|\mathbf{W}_j\|_2 \cdot |\gamma_j| \tag{3}$$

2.3 Evaluation Metrics

To evaluate model generalization, we monitor the training, validation, and test losses alongside their corresponding accuracies. Furthermore, we introduce the AN index to quantify generalization deterioration and track its dynamic progression under varying pruning ratios. Formally, AN is defined as the mean absolute discrepancy between validation loss and training loss over the final K epochs of the fine-tuning phase. Let $E = 40$ denote the total number of fine-tuning epochs, and $K = 20$ represent the averaging window size over the terminal fine-tuning stage. Mathematically, AN is formulated as:

$$\text{AN} = \frac{1}{K} \sum_{e=E-K+1}^{E} \left| \mathcal{L}_{\text{val}}^{(e)} - \mathcal{L}_{\text{train}}^{(e)} \right|, \tag{4}$$

where $\mathcal{L}_{\text{val}}^{(e)}$ and $\mathcal{L}_{\text{train}}^{(e)}$ denote the validation loss and training loss at epoch e, respectively.

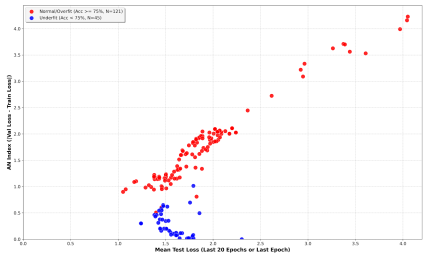


Figure 1: **Scatter plot of the AN index against mean test loss.** Both metrics are averaged over the trailing 20 fine-tuning epochs across 167 ResNet model configurations spanning diverse hyperparameter settings. Blue dots represent the underfitted configurations with train accuracy below 75%, while red dots correspond to normal or overfitted models. Notably, red dots with AN > 2.5 appear exclusively under aggressive pruning without regularization.

Across 167 training runs, excluding underfitted models, AN is positively correlated with test loss (Figure 1): as overfitting worsens, training loss approaches zero while validation loss rises, widening

the gap. Underfitted models cluster near AN $= 0$ because both losses remain high and nearly equal. The core results (non-monotonic generalization, architectural threshold) hold when measured directly by test and validation loss (Appendix, Figure 26).

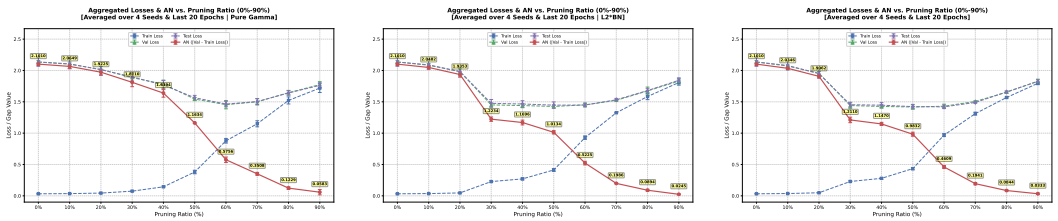
3 RESULTS

3.1 FINDING 1: THE DEEPEST BLOCK CONCENTRATION

Motivation: We observe that in the overparameterized regime, the model's generalization improves as pruning ratio increases before degrading after a critical threshold, and that deeper, parameter-dense channels are preferentially targeted. While the latter aligns with Network Slimming (Liu et al., 2017), the mechanism driving this localized concentration and the non-monotonic generalization trajectory remain unexplained. By tracking the dynamic evolution of BN scale factors (γ) and L_2 norms (Figure 3), we find that **redundant deep channels undergo a dual collapse, and pruning preferentially concentrates in those channels—driving the initial generalization recovery. We term the deepest active block, where this structural suppression is most pronounced, the *"noise sink"* of the network. The generalization improvement continues until the noise sink is exhausted and genuine feature-extracting capacity begins to be damaged.**

Experiment results: On ResNet-18/CIFAR-10, all three criteria unanimously target Block 4 below the 53.33% threshold (Table 3). BN scale factors and L_2 norms undergo a severe dual collapse exclusively in Block 4 during pre-training (Figure 3), marking it as functionally redundant (Li et al., 2017; Liu et al., 2017). Block removal (Figure 4) confirms Block 4 is not merely redundant: dropping it yields an immediate sharp improvement in generalization (AN drop, test/val loss decrease, training loss spike), suggesting it carries active overfitting-related capacity.

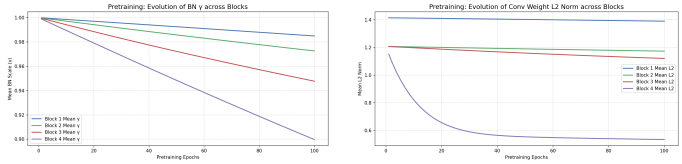
The same dual collapse and Block-4-first pruning pattern is reproduced on ResNet-18/CIFAR-100 and ResNet-34/CIFAR-100 (Appendix Figures 19, 21, 20, 22), confirming that the "noise sink" behavior is not specific to ResNet-18/CIFAR-10.



(a) BN scale factors (γ) (b) $L_2 \times \gamma$ (c) L_2

Figure 2: **Non-monotonic generalization behavior across pruning ratios for three importance criteria on the 10% CIFAR-10 subset.** All three criteria show consistent generalization improvement below the 53.33% architectural threshold, followed by degradation at higher ratios. AN index (red), test loss (purple), and validation loss (green) are shown alongside training loss (blue). The non-monotonic trajectory motivates the block-level structural analysis in Finding 1.

Mechanism: A proposed interpretation of why these two phenomena occur simultaneously in deepest blocks. We propose the following interpretation, which is consistent with our empirical observations. When model capacity far exceeds data capacity, the optimization process is driven by "simplicity bias", preferentially learning low-frequency and simple patterns from the data (Arpit et al., 2017). According to the hierarchical distribution of features in deep networks, the extraction of such low-level features is topologically undertaken by shallow channels (Yosinski et al., 2014). Representation similarity analysis independently confirms that later layers develop markedly more task-specific representations than earlier ones (Kornblith et al., 2019), consistent with deeper blocks specializing toward dataset-specific patterns rather than transferable features. When the effective task complexity is low relative to model capacity, foundational features extracted by shallow layers are already sufficient to cover all effective information. Consequently, during this stage, the model



(a) γ evolution during pre-train (b) L_2 evolution during pre-train

Figure 3: **BN scale factor (γ) and L_2-norm trajectories for each ResNet-18 block during pre-train.** In this experiment, we monitored the mean γ and L_2 of all the channels in each blocks, and analyzed their dynamic evolutions during pre-train. Both metrics undergo a severe dual collapse exclusively in Block 4, while L_2-norm in Blocks 1-3 remain stable, and γ in Blocks 1-3 experience collapse far less drastic than that in Block 4. This simultaneous collapse is observed consistently across all four random seeds and identifies Block 4 as the functional "noise sink".

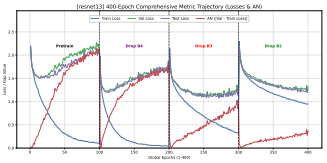


Figure 4: **Generalization metrics following sequential block removal.** Removing Block 4 produces an immediate sharp improvement in generalization (AN drop, test/val loss decrease), confirming that Block 4 carries concentrated overfitting-related capacity rather than being merely redundant.

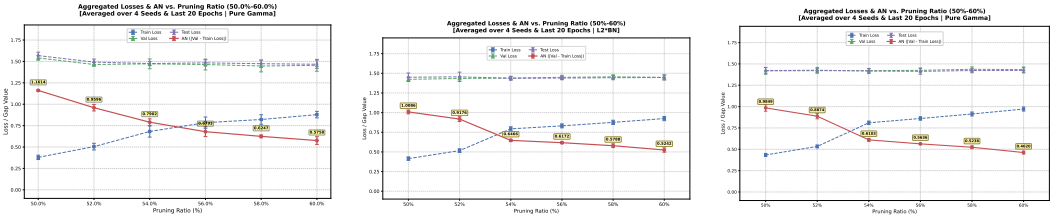
structurally prioritizes shallow channels, while parameter-dense deeper channels remain essentially idle, which can be confirmed by the dual collapse of γ and L_2 in Figure 3.

The deepest block, being topologically closest to the loss, is the first to respond to whatever residual signal the shallower layers do not account for. Under this interpretation, the deepest active block accumulates the residual overfitting-related capacity—observable as near-100% training accuracy together with the dual collapse of γ and L_2 exclusively in Block 4 (Figure 3). Critically, a controlled label-noise experiment (Appendix, Figure 27) shows that Block 4's γ and L_2 collapse is *invariant* to label-noise level, indicating the concentration reflects general overfitting-related capacity rather than a response to label corruption specifically. An optimizer ablation (Appendix, Figure 25) shows L_2 suppression persists under SGD+momentum while gamma collapse is amplified by Adam's adaptive learning rates, indicating the structural signal is optimizer-independent in its weight-norm form.

3.2 Finding 2: Sharp Transition Threshold

Motivation: In Finding 1, we established that Block 4 functions as a "noise sink", concentrating the majority of the network's residual memorization capacity. Consequently, removing it deprives the model's capacity for dataset-specific memorization, which is manifested as a boost in generalization. However, we additionally observe that regardless of the pruning strategy (L_2, γ, or $L_2 \times \gamma$) on ResNet-18, under the CIFAR-10 dataset, the model's generalization capabilities undergo a severe sharp transition as the pruning ratio scales from 50% to 60% (Table 2). This transition is characterized by a precipitous drop in the AN metric, validation loss, and test loss, coupled with a sharp spike in training loss. Therefore, this section investigates the underlying cause of this sharp transition and explores whether it correlates with our observed deepest block concentration. To answer this, we designed a fine-grained probing experiment on ResNet-18 under CIFAR-10 dataset, with a 2% increment of pruning ratio between the 50% and 60% sparsity levels, recording the block-wise distribution of pruned channels at each ratio. Furthermore, we validate our finding on ResNet-34 under CIFAR-10. **The learned importance landscape concentrates pruning pressure in Block 4; when Block 4's channels are exhausted—at the architecturally determined stage-exhaustion boundary of 53.33% (ResNet-18) or 40.68% (ResNet-34)—the model must begin selecting channels from other blocks. Crossing this boundary is necessary but not sufficient for a sharp generalization event: the γ criterion produces a gradual transition, while magnitude-sensitive criteria (L_2, $L_2 \times \gamma$) produce a sharp transition triggered by a single noise-sensitive Block 2 channel that becomes accessible once Block 4 is exhausted.**

Experimental Results and Explanations: Through our fine-grained probing experiments, we found that both the L_2 and $L_2 \times \gamma$ groups exhibit a sharp transition in AN and training loss specifically between 52% and 54% (Figure 5b and Figure 5c, Table 2), whereas the γ group (Figure 5a) displays a much smoother, progressive changing. Furthermore, we deconstructed the micro-



(a) BN scale factors (γ) (b) $L_2 \times \gamma$ (c) L_2

Figure 5: **Fine-grained generalization curves across the 50-60% pruning range at 2% resolution.** The γ criterion (left) exhibits a gradual, continuous transition across the 53.33% architectural threshold. The $L_2 \times \gamma$ (center) and L_2 (right) criteria exhibit a sharp transition between 52% and 54%, triggered by the selection of a noise-sensitive Block 2 channel once Block 4 is exhausted (see Figure 8 in Appendix). This criterion-dependent behavior demonstrates that the 53.33% threshold is necessary but not sufficient for a sharp generalization transition.

Table 1: **Architectural decomposition of internal prunable channel capacities for ResNet-18 and ResNet-34.** In ResNet-18, Block 4 alone holds 53.33% of the total prunable channels. In contrast, ResNet-34 is deeper in its intermediate stages, with Block 3 and Block 4 equally dominating the capacity (40.68% each).

Block Name	Width (C)	Convolution Layers	Total Channels	Prunable Channels	% of Total
ResNet-18 Architecture					
Block 1	64	4	$64 \times 2 = 128$	128	6.67%
Block 2	128	4	$128 \times 2 = 256$	256	13.33%
Block 3	256	4	$256 \times 2 = 512$	512	26.67%
Block 4	512	4	$512 \times 2 = 1024$	1024	**53.33%**
Total	—	**16**	—	**1920**	**100.00%**
ResNet-34 Architecture					
Block 1	64	6	$64 \times 3 = 192$	192	5.08%
Block 2	128	8	$128 \times 4 = 512$	512	13.56%
Block 3	256	12	$256 \times 6 = 1536$	1536	**40.68%**
Block 4	512	6	$512 \times 3 = 1536$	1536	**40.68%**
Total	—	**32**	—	**3776**	**100.00%**

architecture of ResNet-18 and ResNet-34 (He et al., 2016). For ResNet-18, the network consists of 4 layers, each containing 2 basic blocks. From the first to the fourth block, there are 128, 256, 512, and 1024 channels, respectively. Crucially, the channels in the fourth block account for exactly 53.33% of the network's total capacity—situated perfectly between our 52% and 54% probe points. Aligning this structural fact with our channel selection statistics, we observed that prior to the 52% pruning ratio, all three criteria exclusively target Block 4 for pruning. Once the 53.33% boundary is crossed (Block 4 is completely stripped), the model is forced to select channels from outside Block 4, as shown in Figure 8. Therefore, we conclude that the 53.33% is an architecturally-determined threshold which can predict exactly when channels in blocks other than Block 4 become pruning candidates.

However, for ResNet-18, models pruned with the three criteria exhibit two distinct generalization behaviors at this critical boundary, which we explain separately below:

(a) γ Criterion: The transition is gradual. No sharp event occurs at the 53.33% threshold. Prior to reaching 53.33%, AN decreases continuously, training loss increases continuously, and test loss holds stable. After crossing 53.33%, the model strictly selects Block 3 channels as additional pruning candidates, as shown in Figure 8a. This presents a clean picture of progressive noise deprivation: channels in Block 4 are removed incrementally, and each removal reduces the model's memorization capacity without triggering a sudden structural event. The rate of change of AN is largest in the 40%-60% range, which is consistent with the threshold boundary.

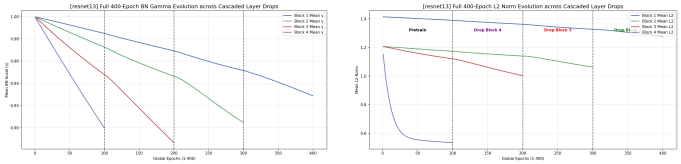
(b) L_2 and $L_2 \times \gamma$ Criteria: A sharp transition occurs at the threshold. We observe that after crossing the 53.33% boundary, in addition to the supplementary pruning of Block 3 channels, exactly one channel belonging to Block 2 is also pruned in Figure 8b and 8c. This constitutes the sole distinction in channel selection between these two criteria and the γ criterion, yet it triggers a sharp increase in training loss and AN. This indicates that noise-absorbing channels, while primarily concentrated in Block 4, are not exclusively located there. Shallow-block noise-absorbing channels appear significantly more sensitive to removal, as evidenced by the disproportionate generalization impact of this single Block 2 channel compared to the gradual response observed during the extensive removal of Block 4.

For the experiments on ResNet-34, a very similar pattern emerges. On the CIFAR-10 dataset, under the L_2-based pruning criterion, channels in Block 4 are exclusively removed. Just before crossing the critical threshold of 40.68% determined by the ResNet-34 architecture, the model's generalization undergoes a sharp transition at a pruning ratio between 40% and 50% (Figure 9 in Appendix), which mainly happens between 40% and 42%, where the model crosses the architecturally determined threshold of 40.68% (Figure 11 in Appendix). In contrast, under the γ-based pruning criterion, channels from both Block 3 and Block 4 are removed simultaneously—though Block 4 still dominates the quantity—and the generalization transition between the 40% and 50% ratios is not pronounced. This once again demonstrates that the severity of the sharp transition is inherently criteria-dependent. However, we also observed that in ResNet-34, L_2-based pruning consistently yields a globally lower AN metric compared to the γ criterion (Figure 9 in Appendix). This occurs because once all channels in Block 4 are exhausted, the L_2 criterion begins selecting channels from both Block 2 and Block 3 simultaneously (Figure 10 in Appendix). In contrast, prior to reaching an 80% pruning ratio, the γ criterion exclusively selects additional channels from Block 3. This further corroborates our earlier proposition that shallower noise-absorbing channels exhibit a higher sensitivity to removal.

3.3 Finding 3: Deepest-Stage Migration

Motivation: Having established that Block 4 acts as a "noise sink" within the overparameterized regime, we next need to determine whether this "noise sink" is a static property fixed to the deepest block, or a dynamic role dictated by topological depth. To answer this question, we designed a sequential block removal experiment (Figure 4 and Figure 6). We first train the original model for 100 epochs, then remove all channels of the current deepest layer and continue training for another 100 epochs. This procedure is repeated iteratively until the model is reduced to only all the channels in Block 1, after which we train for a final 100 epochs and record the ultimate results. Throughout this 400-epoch process, we continuously track the dynamic evolution of the γ and L_2 metrics. **Ultimately, we discovered that regularization pressure does indeed dynamically migrate to the second-deepest block following the removal of the entire deepest block. In other words, functioning as a "noise sink" is a characteristic of the "deepest active layer" in our experiment.**

Experiment Results: Through the block removal experiment on ResNet-18 and ResNet-34 (Figure 4,Figure 12 in Appendix), we discovered that following the removal of the entire current deepest block, the γ and L_2 metrics of the "active deepest layer" (formerly the second-deepest) consistently exhibit a dual collapse (Figure 6,Figure 13 in Appendix). Although the intensity of this collapse is slightly less severe than in the initial pre-training phase, it demonstrates that the structural differentiation pattern—with the new deepest block developing suppressed γ/L_2 relative to shallower blocks—re-emerges after each removal. Furthermore, during subsequent training phases after dropping Block 4 and Block 3, test loss, validation loss, and AN continuously increase while training loss declines (Figure 4, Figure 12 in Appendix), consistent with the new deepest block accumulating overfitting-related capacity. In summary, both the structural differentiation pattern and the concentrated overfitting-related capacity shift to the next-deepest active block following each removal, suggesting the "noise sink" role is a property of topological depth rather than a fixed characteristic of any specific block. A parallel block-removal experiment on VGG-16 (Appendix, Figure 14) confirms that this migration is not ResNet-specific: after removing VGG-16's deepest block (B5), the suppression statistic of the new deepest block (B4) doubles from $S_\gamma = 0.073$ to 0.144.



(a) BN scale factors (γ) (b) L_2

Figure 6: **Dynamic evolution of BN scale factor (γ) and L_2 norm during the sequential block removal experiment.** Following the removal of Block 4 at epoch 100, the γ and L_2 metrics of Block 3 (the new deepest active block) undergo a dual collapse mirroring the original Block 4 behavior. This migration of collapse patterns to the next-deepest block repeats after Block 3 is removed at epoch 200, confirming that the "noise sink" role is a dynamic property of topological depth rather than a fixed characteristic of any specific block.

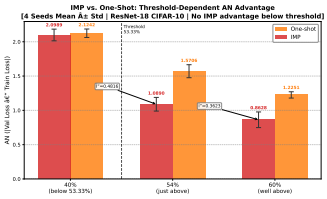


Figure 7: **IMP vs One-Shot Pruning.** In the experiments using ResNet-18 on 10% CIFAR-10, the AN metrics of IMP and one-shot pruning were compared across different pruning ratios, demonstrating that the advantage of IMP is primarily manifested after crossing the threshold.

Immediately following each block removal, the AN index drops sharply and training loss spikes, confirming that the deepest block concentrates memorization capacity. The highest test accuracy (65.2%) across all ResNet-18 runs occurred during training after the removal of Block 3, consistent with sequential noise sink deprivation.

4 Discussion

4.1 Microscopic Mechanism of the Lottery Ticket Hypothesis

Interpretation of the Generalization Improvement of the Lottery Ticket Hypothesis via "noise sink": Discarding "Losing Tickets" also matters. The Lottery Ticket Hypothesis (LTH) (Frankle & Carbin, 2019) posits that pruning improves model generalization by uncovering a well-initialized, sparse subnetwork (the "winning ticket") while discarding the remaining 'useless' parameters. However, our empirical findings in the overparameterized regime offer a complementary, micro-architectural perspective: the removed channels—typically regarded merely as "losing tickets"—are not passively redundant but carry concentrated overfitting-related capacity. Block removal experiments (Figure 4) show that stripping away these deep redundant weights produces a consistent recovery in generalization. Under this view, the generalization boost from pruning in the overparameterized regime is not solely attributable to the preservation of the winning ticket; it is also attributable to the structural removal of the concentrated overfitting-related capacity from the deepest stage. This structural specificity is confirmed by a randomized pruning control (Appendix, Figure 29): under random channel selection the removal load is approximately uniform across all blocks, whereas under BN γ ranking Block 4 absorbs 81–100% of the removal—demonstrating that preferential deepest-stage targeting reflects learned weight structure, not the pruning procedure itself. Furthermore, a controlled label-noise experiment (Appendix, Figure 27) shows that Block 4 γ and L_2 are invariant to label-noise level, confirming that the observed concentration reflects overfitting-related memorization capacity rather than a specific response to label corruption.

Structural Perspective of Iterative Magnitude Pruning via the Migration Phenomenon IMP (Frankle & Carbin, 2019) interleaves pruning with retraining and outperforms one-shot pruning with a matched training budget (6 rounds $\times$ 40 fine-tuning epochs = 240 total fine-tuning epochs each). We propose this advantage is consistent with noise sink migration: iterative retraining allows memorization capacity to fully migrate to the second-deepest block before each pruning step, so each IMP iteration targets the current noise sink accurately. One-shot pruning beyond the architectural threshold instead indiscriminately removes feature-extracting channels. Empirically, IMP and one-shot pruning achieve comparable AN at 40% (below threshold), while IMP's AN drops significantly below one-shot pruning after crossing the 53.33% threshold (Figure 7; Appendix Figure 15).

4.2 Generality Across Dataset Scales and Architectures Beyond ResNet

To evaluate the robustness of our findings, we deployed the full CIFAR-10 dataset on ResNet-18 and conducted experiments across pruning ratios from 20% to 60%. We observe that the dual collapse of both γ and L_2 norms occurs during pre-train on the full dataset as well, and channel selection from 20% to 60% still predominantly targets Block 4 (see Table 4 in Appendix). On full CIFAR-10, test accuracy remains flat ($74.3\% \pm 0.4\%$ to $73.4\% \pm 0.6\%$) across all pruning ratios while AN improves monotonically (Appendix Table 5), indicating that the non-monotonic accuracy trajectory is specific to the overparameterized 10% regime. A controlled data-fraction sweep (Appendix, Figure 28) reveals that Block 4 suppression strengthens monotonically with training data: defining $S_\gamma = 1 - \bar{\gamma}_{B4}/\bar{\gamma}_{B1:B3}$ as the relative suppression of Block 4 mean BN scale, S_γ increases from 0.038 at 5% to 0.434 at 100% of full CIFAR-10. This indicates that depth-dependent structural differentiation is a property of network convergence—a well-trained model develops stronger Block 4 suppression as it processes more training signal—rather than a pathology exclusive to extreme data scarcity. As a preliminary test of architectural generality beyond ResNet, we applied the gamma criterion to VGG-16 under identical experimental conditions (four seeds, from 42 to 45). VGG-16's five convolutional block groups contain 128, 256, 768, 1536, and 1024 prunable internal channels respectively, giving a total of 3,712 channels (see Appendix Table 7 for full decomposition). We observe that gamma collapse is shared between Blocks 4 and 5 — the two deepest blocks — with Block 5 dominant. Channel selection from 10% pruning onwards targets exclusively Blocks 4 and 5 (see Appendix Table 6), with Block 3 entering selection at approximately 70%, after Blocks 4 and 5 are nearly exhausted. This is consistent with the principle that collapsed deep stages are preferentially targeted: when two blocks share the noise-sink role (as in VGG-16's deeper geometry), both are exhausted before shallower blocks are touched. We present this as preliminary evidence that the depth-dependent structural concentration pattern extends beyond ResNet models. Additionally, a supplementary experiment on TinyImageNet (64×64, 200 classes; Appendix, Figure 24) confirms the dual collapse at higher resolution: $S_\gamma = 0.123$ and $S_{L2} = 0.224$ at epoch 100 under identical training conditions, showing the phenomenon is not specific to 32×32 imagery.

5 Conclusion

This paper finds that, in the overparameterized regime, overfitting-related structure disproportionately concentrates in the deepest active block—which we term the "noise sink"—and that this depth-dependent concentration migrates to the next-deepest block when the current deepest is removed. Controlled experiments show the structural signal is present under SGD (L_2 suppression) and independent of label noise, while adaptive optimizers amplify the gamma collapse. We identify an architecturally determined pruning threshold (53.33% for ResNet-18; 40.68% for ResNet-34) beyond which generalization undergoes a sharp criterion-dependent transition, and a structural observation specific to the overparameterized regime: test accuracy peaks at ratios below the threshold while test loss continues to improve slightly through it.

Limitations. All primary experiments use CIFAR-scale imagery; a TinyImageNet supplementary (Appendix, Figure 24) extends to 64×64, but ImageNet-scale validation remains open. The dual collapse is substantially more pronounced under Adam than SGD+momentum (Appendix, Figure 25): L_2 suppression persists under both, but gamma collapse is Adam-specific. The proposed mechanism is a qualitative interpretation, not a formally proved theorem. The IMP advantage is *consistent with* noise sink migration but does not constitute direct causal proof. The AN index reliably tracks generalization in the noise-deprivation and threshold-crossing phases, but understates deterioration at extreme pruning ratios ($\geq 70\%$) where both training and test loss rise simultaneously: the train–test gap narrows even as the model degrades, so AN should not be interpreted as a generalization proxy beyond the capacity-destruction boundary. The higher sensitivity of shallow-block channels to removal (Finding 2) may reflect unequal regularization pressure: with fewer channels per block, each shallow channel absorbs disproportionately more optimization signal, making its removal more disruptive; this remains a hypothesis not yet formally tested.

Reproducibility Statement

All experiments use four random seeds (42–45) with `torch.backends.cudnn.deterministic = True`. Hyperparameters, dataset splits, and pruning protocols are fully specified in Section 2. The architecturally determined thresholds (53.33% for ResNet-18; 40.68% for ResNet-34) follow directly from the block channel-capacity fractions in Table 1. Code for all experiments, including seed-level CSV outputs and the figure-generation script, is available at `https://anonymous.4open.science/r/depth-structure-pruning-CB93/`.

Ethics Statement

This work studies structural properties of neural network pruning using publicly available image classification benchmarks (CIFAR-10, CIFAR-100, TinyImageNet). No human subjects were involved, no new datasets were collected, and no personally identifiable information was used. The findings carry no direct risks of misuse or harm.

AI Use Statement

The authors used LLM assistance (Claude, Anthropic) for grammar refinement, text restructuring, and figure caption editing during manuscript preparation. All scientific content, experimental design, data analysis, and conclusions are solely the work of the authors. The LLM was not used to generate or interpret experimental results.

References

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, and Yoshua Bengio. A closer look at memorization in deep networks. In *International Conference on Machine Learning (ICML)*, pp. 233–242. PMLR, 2017.

Brian R. Bartoldson, Ari S. Morcos, Adrian Barbu, and Gordon Erlebacher. The generalization-stability tradeoff in neural network pruning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJl-b3RcF7`.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2 edition, 2009.

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 (CVPR)*, pp. 770–778, 2016.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

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

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In *Advances in neural information processing systems (NeurIPS)*, pp. 598–605. 1989.

Hao Li, Aseem Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *International Conference on Learning Representations (ICLR)*, 2017.

Jingling Li, Yanchao Sun, Jiahao Su, Taiji Suzuki, and Furong Huang. Understanding generalization in deep learning via tensor methods. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 504–515. PMLR, 2020.

Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pp. 2736–2744, 2017.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations (ICLR)*, 2019.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, 2014.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

A Related Works

A.1 Structural Pruning

Structural pruning is a coarse-grained, hardware-friendly technique utilized to compress neural networks and accelerate inference while preserving predictive accuracy (Li et al., 2017). The L_2-norm and Batch Normalization (BN) scaling factor γ are commonly employed to evaluate the importance of structural components, allowing for the removal of channels that contribute minimally to the model (Liu et al., 2017; Li et al., 2017). Conversely, (Liu et al., 2019) posited that the pruned architecture itself, rather than the inherited weights, is the critical factor in achieving high efficiency, framing pruning as an implicit architecture search mechanism. In our study, we do not employ pruning to improve inference efficiency; instead, we utilize it as an analytical tool to explore the model's micro-architecture. Specifically, we investigate the functional asymmetry across network blocks using the L_2-norm and BN γ factors, revealing that in ResNet-18, Block 4 is highly redundant and acts as "noise sink".

A.2 Generalization in Compressed Networks

Deep neural networks typically possess massive parameter counts, leading to substantial redundancy. Models often exploit this excess capacity to memorize dataset labels as a shortcut, thereby artificially inflating training accuracy (Zhang et al., 2017). However, this memorization fails on test sets with uncorrelated labels, ultimately resulting in overfitting—a pronounced generalization gap between training and test errors (Hastie et al., 2009)—which is exacerbated in data-scarce regimes. Existing research indicates a fundamental connection between a network's compressibility and its generalization capabilities; compression methods like pruning not only reduce parameter counts and computational overhead but also influence generalization dynamics (Bartoldson et al., 2020; Li et al., 2020). Our findings reveal an architecture-dependent pruning threshold. Crossing this threshold induces a sharp transition in model generalization, maximally stripping the network of its capacity to memorize noisy labels.

A.3 Lottery Ticket Hypothesis

Frankle & Carbin (2019) proposed that within a randomly initialized, dense neural network, there exist "winning tickets"—sparse subnetworks that, when trained in isolation with their original initialization, can achieve performance comparable to the original unpruned network. This hypothesis

explains why a trained, overparameterized network can be heavily pruned while maintaining or recovering its original performance. Our analysis reveals that noise within these networks is primarily concentrated in redundant deep layers, and pruning inherently prioritizes the removal of these deep channels. Consequently, improving model generalization requires not only identifying and preserving well-initialized subnetworks conducive to training but also actively eliminating "losing tickets". The Lottery Ticket Hypothesis (LTH) employs Iterative Magnitude Pruning (IMP) to uncover these sparse, trainable subnetworks, noting that IMP yields better generalization improvements than one-shot pruning. We show that IMP outperforms one-shot pruning in a manner consistent with the view that iterative retraining affords the network sufficient time to complete the upward migration of regularization pressure and memorization capacity.

B Supporting Material — Findings 1 & 2

Tables and figures below provide detailed per-criterion and per-block statistics supporting Findings 1 and 2 (channel concentration and the architecturally determined threshold).

Table 2: **Summary of training dynamics and generalization milestones across all three pruning criteria.** Data represents the mean ± std of the final 20 epochs across 4 random seeds (seeds 42–45). The 0% baseline serves as the global anchor for the unpruned model. Across these three criteria, the train accuracy experienced a sharp drop, falling from around 90% to approximately 70% from 50% to 60% pruning ratio, which means the model's memorization capacity is substantially deprived. From 52% to 54% pruning ratio, during which the model across the 53.33% threshold, the generalization experienced sharp transition that the train loss increased drastically while test loss, train accuracy and AN index sharply declined.

Criterion	Ratio	Train Loss	Test Loss	Train Acc (%)	Test Acc (%)	AN Index
Baseline	**0%**	**0.0325 ± 0.0035**	**2.1364 ± 0.0806**	**99.21 ± 0.11**	**52.92 ± 0.96**	**2.10 ± 0.03**
L_2 **Norm**	40%	0.2791 ± 0.0157	1.4464 ± 0.0467	95.00 ± 0.32	53.60 ± 1.90	1.15 ± 0.03
	50%	0.4339 ± 0.0211	1.4232 ± 0.0411	89.99 ± 0.58	53.26 ± 1.86	0.98 ± 0.04
	52%	0.5333 ± 0.0200	1.4260 ± 0.0390	86.48 ± 0.60	52.80 ± 1.28	0.89 ± 0.04
	54%	0.8104 ± 0.0229	1.4169 ± 0.0345	78.03 ± 1.29	51.67 ± 1.82	0.61 ± 0.02
	60%	0.9688 ± 0.0270	1.4196 ± 0.0317	70.62 ± 1.44	51.36 ± 1.61	0.46 ± 0.03
BN γ	40%	0.1438 ± 0.0102	1.7672 ± 0.0849	96.54 ± 0.35	52.88 ± 2.13	1.64 ± 0.08
	50%	0.3800 ± 0.0325	1.5655 ± 0.0428	90.41 ± 1.06	52.67 ± 2.13	1.16 ± 0.01
	52%	0.5048 ± 0.0448	1.4925 ± 0.0408	87.10 ± 1.36	52.21 ± 1.52	0.96 ± 0.04
	54%	0.6817 ± 0.0756	1.4754 ± 0.0285	80.99 ± 2.39	51.87 ± 1.00	0.79 ± 0.05
	60%	0.8786 ± 0.0425	1.4642 ± 0.0510	75.04 ± 1.41	51.39 ± 1.57	0.58 ± 0.05
Joint $L_2 \times \gamma$	40%	0.2701 ± 0.0176	1.4649 ± 0.0606	95.27 ± 0.37	53.72 ± 1.79	1.17 ± 0.04
	50%	0.4128 ± 0.0257	1.4469 ± 0.0590	90.63 ± 0.74	53.34 ± 2.49	1.01 ± 0.03
	52%	0.5163 ± 0.0272	1.4551 ± 0.0685	87.01 ± 0.95	52.44 ± 2.30	0.92 ± 0.04
	54%	0.7932 ± 0.0307	1.4365 ± 0.0292	78.87 ± 1.43	52.26 ± 1.73	0.65 ± 0.02
	60%	0.9281 ± 0.0324	1.4484 ± 0.0312	72.27 ± 1.66	50.84 ± 0.92	0.52 ± 0.03

C Supporting Material — Finding 3: Deepest-Stage Migration

Sequential block-removal experiments confirm that the noise-sink role migrates to the next-deepest active block after removal; the pattern holds on both ResNet and VGG architectures.

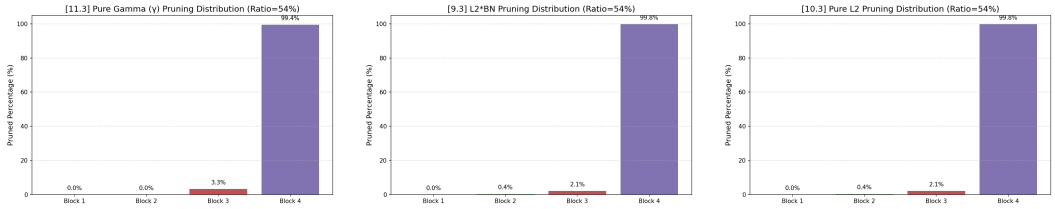
C.1 ResNet-34 Block Removal

C.2 VGG-16 Block Removal

Three sequential 100-epoch phases confirm that the suppression statistic doubles for the new deepest block after removal on VGG-16, replicating the migration pattern on a non-ResNet architecture.

Table 3: **Layer-wise pruning statistics for ResNet-18 across three pruning criteria under global pruning ratios (Averaged over 4 seeds).** Total prunable channels = 1920. Each block retains a minimum of two channels—one per Basic Block—to enforce a safety guard that prevents zero-channel tensor dimensions from causing forward-pass runtime errors during extreme pruning. Only channels in Block 4 are pruned before crossing the 53.33% threshold across three pruning criteria. Beyond the threshold, the model with pure γ criterion merely selects channels in Block 3 as additional pruning candidates, while the one with L_2 and $L_2 \times \gamma$ selects channels in Block 2 and 3, which means the channel-selection beyond threshold is criteria-dependent.

Prune Criterion	Ratio	Total Pruned	Block 1 (C=128)	Block 2 (C=256)	Block 3 (C=512)	Block 4 (C=1024)
	0.20	384 / 1920	0	0	0	384 (37.50%)
	0.30	576 / 1920	0	0	0	575 (56.15%)
	0.40	768 / 1920	0	0	0	767 (74.90%)
L_2-norm	0.50	960 / 1920	0	0	0	959 (93.65%)
	0.52	998 / 1920	0	0	0	997 (97.36%)
	0.54	1036 / 1920	0	2.2±1.6 (0.86%±0.62%)	9.8±1.6 (1.91%±0.31%)	1022 (99.80%)
	0.60	1152 / 1920	0	26.5±2.3 (10.35%±0.90%)	101.5±2.3 (19.82%±0.45%)	1022 (99.80%)
	0.20	384 / 1920	0	0	0	384 (37.50%)
	0.30	576 / 1920	0	0	0	576 (56.25%)
	0.40	768 / 1920	0	0	0	768 (75.00%)
BN γ	0.50	960 / 1920	0	0	0	960 (93.75%)
	0.52	998 / 1920	0	0	1.2±0.4 (0.23%±0.08%)	996.5±0.9 (97.31%±0.09%)
	0.54	1036 / 1920	0	0	17.0±1.2 (3.32%±0.23%)	1018.0±1.2 (99.41%±0.12%)
	0.60	1152 / 1920	0	0	128 (25.00%)	1022 (99.80%)
	0.20	384 / 1920	0	0	0	384 (37.50%)
	0.30	576 / 1920	0	0	0	575 (56.15%)
	0.40	768 / 1920	0	0	0	767 (74.90%)
Joint $L_2 \times \gamma$	0.50	960 / 1920	0	0	0	959 (93.65%)
	0.52	998 / 1920	0	0	0	997 (97.36%)
	0.54	1036 / 1920	0	0.5±0.5 (0.20%±0.20%)	11.5±0.5 (2.25%±0.10%)	1022 (99.80%)
	0.60	1152 / 1920	0	8.2±2.9 (3.20%±1.13%)	119.8±2.9 (23.40%±0.57%)	1022 (99.80%)



(a) BN scale factors (γ) (b) $L_2 \times \gamma$ (c) L_2

Figure 8: **Block-wise channel selection distribution at 54% pruning ratio for all three criteria.** Under the γ criterion (left), pruned channels originate exclusively from Blocks 3 and 4, consistent with a clean architectural boundary crossing. Under $L_2 \times \gamma$ (center) and L_2 (right), exactly one Block 2 channel with an anomalously low magnitude score is additionally selected, triggering the sharp generalization transition observed in Figure 5. This confirms that noise-absorbing channels, while concentrated in Block 4, are not exclusive to it.

D IMP VS ONE-SHOT: BLOCK-WISE CHANNEL CASCADE

IMP performs iterative retraining between removal rounds; the block-wise cascade below shows how exclusive Block 4 targeting persists across rounds, consistent with the view that iterative retraining affords the network sufficient time to complete the upward migration of memorization capacity.

E FOUR-PHASE MODEL OF STRUCTURED PRUNING IN THE OVERPARAMETERIZED REGIME

Integrating our comprehensive findings, we delineate the four-phase model that network generalization undergoes as the pruning ratio increases in the overparameterized regime:

1. **Continuous noise deprivation:** Only Block 4 channels are removed. Train loss rises incrementally as noise memorization capacity decreases. Test loss improves or holds stable. The AN index decreases. No structural damage occurs.
2. **Block-level threshold crossing:** At 53.33%, Block 4 is exhausted and pruning candidates emerge from shallower blocks. The transition is gradual under pure γ; sharp under

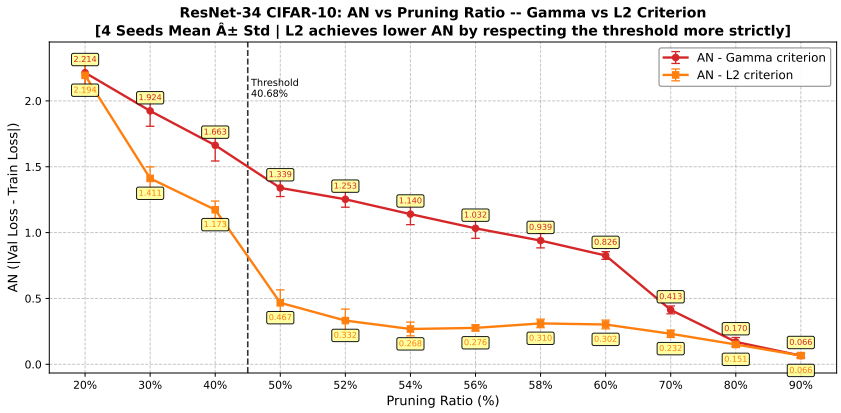


Figure 9: γ **vs** L_2 **criterion upon AN index.** The L_2 criterion undergoes more pronounced transition while the transition of γ criterion is smoother. Meanwhile, the AN value of L_2 is globally less than the one of γ. Both of which validate that the architecturally determined threshold is criteria-dependent.

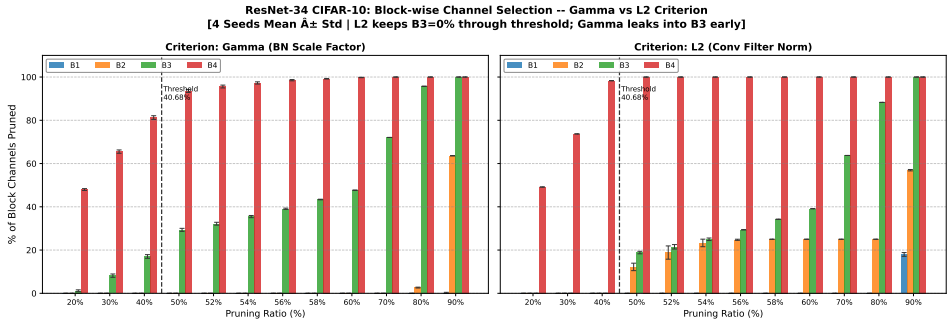


Figure 10: γ **vs** L_2 **criterion upon channel selection.** Before crossing the threshold, the L_2 criterion exclusively targets channels from block 4, while γ targets channels from both 3 and 4. After crossing the threshold, where block 4 channels are exhausted, the L_2 targets channels from both block 2 and 3 for additional selection, while γ targets block 3 channels only, before reaching 80% pruning ratio.

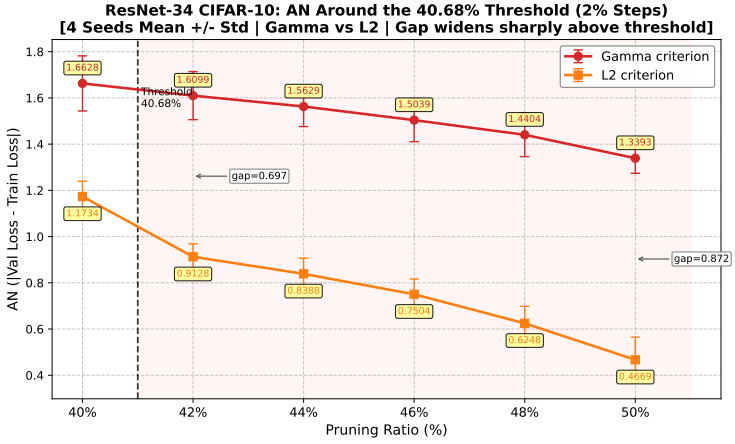


Figure 11: **Fine-grained generalization curves across the 50–60% pruning range at 2% resolution on ResNet-34.** While pruning ratio crossing the architecturally determined threshold 40.68%, the AN index of L_2 criterion is under sharp transition. In contrast, the change of AN index of γ criterion is smoother. This represents the severity of sharp transition is criteria-dependent and decided by the sensitivity to removal of chosen noise-absorbing channels.

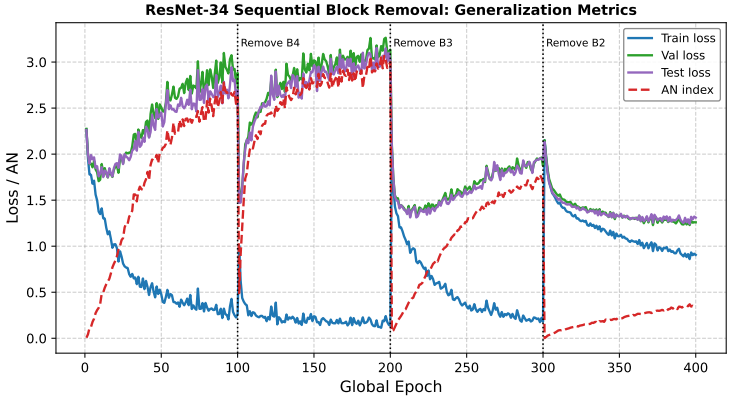
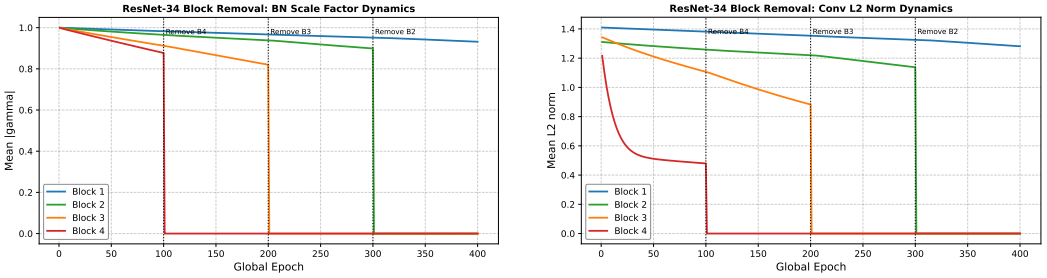


Figure 12: **Generalization metrics following sequential block removal on ResNet-34.** Same as ResNet-18, removing Block 4 produces an immediate sharp improvement in generalization (AN drop, test/val loss decrease) on ResNet-34, confirming that Block 4 contributes substantially to dataset memorization rather than merely being redundant.



(a) γ (b) L_2

Figure 13: **Dynamic evolution of BN scale factor (γ) and L_2-norm during the sequential block removal experiment on ResNet-34.** On ResNet-34, same with the situation on ResNet-18, following the removal of Block 4 at epoch 100, the γ and L_2 metrics of Block 3 (the new deepest active block) undergo a dual collapse mirroring the original Block 4 behavior.

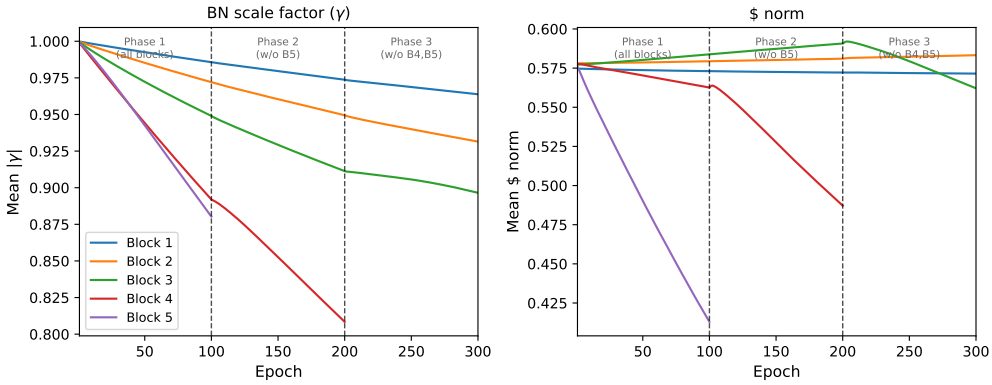


Figure 14: **Structural dynamics during sequential block removal on VGG-16.** Three 100-epoch training phases: Phase 1 uses all five blocks; Phase 2 removes Block 5 (B4 becomes deepest); Phase 3 removes Block 4 (B3 becomes deepest). Dashed vertical lines mark each removal. *Left:* mean BN γ per active block. *Right:* mean L_2 norm. In Phase 1, Block 5 (deepest) shows the lowest γ ($S_\gamma = 0.073$). After its removal, Block 4 becomes the deepest active block and its suppression doubles ($S_\gamma = 0.144$). After Block 4 is removed, Block 3 inherits the deepest role but shows only modest suppression ($S_\gamma = 0.054$), consistent with weaker overparameterization as the network shrinks. Test accuracy improves across phases: $55.2\% \rightarrow 60.5\% \rightarrow 62.9\%$. This confirms that deepest-stage migration is not specific to ResNet architectures.

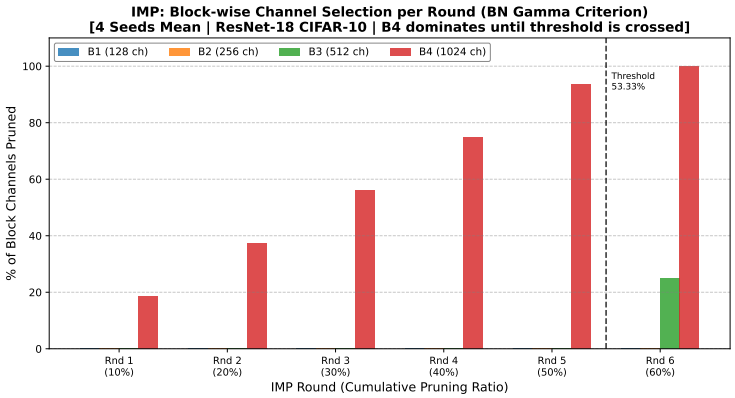


Figure 15: **IMP: Block-wise channel selection per round.** On ResNet-18, channels in Block 4 are exclusively targeted before crossing the 53.33% threshold.

magnitude-sensitive criteria due to the presence of noise-sensitive channels in Block 2. The character of the phase boundary is criterion-dependent.

3. **Generalization optimum:** A pruning ratio exists at which generalization reaches its best observed value. In our experiments, test accuracy peaks at 40% pruning for all three criteria—within the continuous noise deprivation phase, before the architectural threshold is crossed. Test loss reaches its minimum at approximately 54% for L_2 and Joint criteria, and at 60% for the γ criterion. The divergence between accuracy-optimal and loss-optimal pruning ratios is consistent with the coarser resolution of classification accuracy on a small test set: test loss captures improvements in prediction confidence that accuracy cannot resolve. Practitioners should choose the optimal pruning ratio based on their evaluation objective.
4. **Capacity destruction:** Feature-extracting channels are removed. Both train and test loss rise simultaneously. This is not noise deprivation—it is structural damage. The network can no longer perform inference correctly.

F Generality Across Architectures and Datasets

The following subsections extend the structural findings to additional datasets and architectures, confirming that the dual collapse and deepest-stage targeting are not specific to the primary (ResNet-18, 10% CIFAR-10) experimental setting.

F.1 Full CIFAR-10 (ResNet-18)

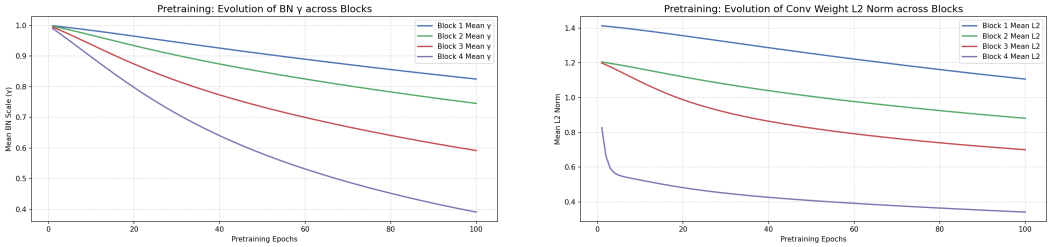
Full CIFAR-10 (45k/5k split) confirms that the dual collapse and Block 4 targeting persist even when the test-accuracy trajectory is flat, indicating that the non-monotonic accuracy curve is specific to the overparameterized regime.

Table 4: **Layer-wise pruning statistics for ResNet-18 under different global BN γ pruning ratios (Seed 42).** Total prunable channels = 1920. Each block retains a minimum of two channels—one per Basic Block—to enforce a safety guard that prevents zero-channel tensor dimensions from causing forward-pass runtime errors during extreme pruning. The model mostly targets Block 4 for pruning.

Prune Ratio	Total Pruned	Block 1 (C=128)	Block 2 (C=256)	Block 3 (C=512)	Block 4 (C=1024)
0.20	384 / 1920	0	0	0	384 (37.50%)
0.30	576 / 1920	0	0	8 (1.56%)	568 (55.47%)
0.40	768 / 1920	0	0	21 (4.10%)	747 (72.95%)
0.50	960 / 1920	0	0	56 (10.94%)	904 (88.28%)
0.52	998 / 1920	0	0	68 (13.28%)	930 (90.82%)
0.54	1036 / 1920	0	0	90 (17.58%)	946 (92.38%)
0.56	1075 / 1920	0	0	109 (21.29%)	966 (94.34%)
0.60	1152 / 1920	0	0	155 (30.27%)	996 (97.27%)

Table 5: **Pruning experiment results on full CIFAR-10 (4-seed mean±std, BN γ criterion, ResNet-18).** Test accuracy is flat across all pruning ratios (less than 1% variation), while AN improves monotonically. Block 4 remains the predominant pruning target throughout.

Prune Ratio	Test Acc (%)	AN Index	Block 4 Pruned / 1024
0.20	74.31 ± 0.42	1.2017 ± 0.0406	383 (37.4%)
0.30	74.36 ± 0.39	1.1866 ± 0.0389	572 (55.9%)
0.40	74.06 ± 0.52	1.1665 ± 0.0435	753 (73.5%)
0.50	74.10 ± 0.27	1.0689 ± 0.0135	913 (89.2%)
0.52	73.92 ± 0.39	1.0141 ± 0.0158	938 (91.6%)
0.54	73.90 ± 0.32	0.9671 ± 0.0168	960 (93.8%)
0.56	73.92 ± 0.41	0.9220 ± 0.0138	979 (95.6%)
0.58	73.71 ± 0.47	0.8938 ± 0.0161	994 (97.1%)
0.60	73.38 ± 0.58	0.8561 ± 0.0235	1006 (98.2%)



(a) γ (b) L_2

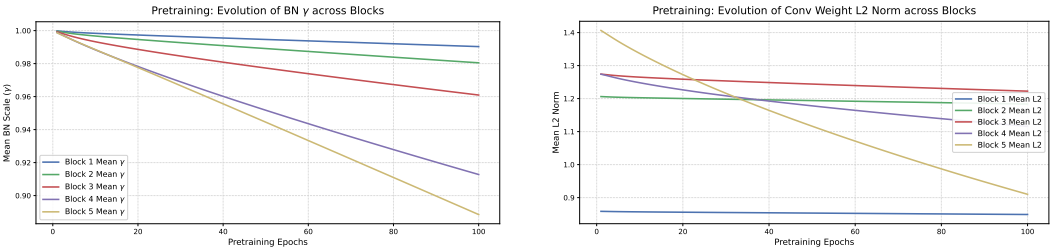
Figure 16: **BN scale factor (γ) and L_2-norm trajectories for each ResNet-18 block during pretrain on full CIFAR-10 dataset.** Both metrics undergo a severe dual collapse exclusively in Block 4 with the lowest values among four blocks, while L_2-norm and γ in Blocks 1–3 experience a mild collapse influenced by weight decay. This confirms that Block 4 remains the noise sink on full CIFAR-10.

Table 6: **Layer-wise pruning statistics (mean $\pm$ std) for VGG-16 under different global BN γ pruning ratios across multiple seeds.** Total prunable channels = 3712. At least one channel is retained in each convolutional layer. Channel selection from 10% pruning onwards targets exclusively Blocks 4 and 5, with Block 3 entering selection at approximately 70% after Blocks 4 and 5 are nearly exhausted.

Prune Ratio	Total Pruned	Block 1 (C=128)	Block 2 (C=256)	Block 3 (C=768)	Block 4 (C=1536)	Block 5 (C=1024)
0.10	371 / 3712	0	0	0	9.5 ± 2.4 ($0.62\% \pm 0.16\%$)	361.5 ± 2.4 ($35.30\% \pm 0.23\%$)
0.20	742 / 3712	0	0	0	71.2 ± 11.4 ($4.64\% \pm 0.74\%$)	670.8 ± 11.4 ($65.50\% \pm 1.11\%$)
0.30	1113 / 3712	0	0	0	246.5 ± 12.9 ($16.05\% \pm 0.84\%$)	866.2 ± 13.4 ($84.59\% \pm 1.31\%$)
0.40	1484 / 3712	0	0	0	522.2 ± 11.1 ($34.00\% \pm 0.73\%$)	961.8 ± 11.1 ($93.92\% \pm 1.09\%$)
0.50	1856 / 3712	0	0	0	858.5 ± 7.2 ($55.89\% \pm 0.47\%$)	997.5 ± 7.2 ($97.41\% \pm 0.70\%$)
0.60	2226 / 3712	0	0	0	1209.0 ± 3.7 ($78.71\% \pm 0.25\%$)	1017.5 ± 3.7 ($99.36\% \pm 0.36\%$)
0.70	2595 / 3712	0	0	46.0 ± 2.4 ($5.99\% \pm 0.32\%$)	1527.0 ± 2.2 ($99.41\% \pm 0.14\%$)	1022 (99.80%)
0.80	2964 / 3712	0	0.2 ± 0.5 ($0.10\% \pm 0.19\%$)	409.5 ± 0.6 ($53.33\% \pm 0.08\%$)	1532.8 ± 0.5 ($99.78\% \pm 0.03\%$)	1022 (99.80%)
0.90	3334 / 3712	0	36.0 ± 2.9 ($14.06\% \pm 1.15\%$)	743.5 ± 2.5 ($96.81\% \pm 0.33\%$)	1533 (99.80%)	1022 (99.80%)

F.2 VGG-16

VGG-16 pruning statistics confirm that the deepest-stage targeting principle extends to a non-ResNet architecture with five blocks and a different capacity distribution.



(a) γ (b) L_2

Figure 17: **BN scale factor (γ) and L_2-norm trajectories for each VGG-16 block during pretrain.** For γ, both Block 4 and Block 5 undergo severe collapse, with Block 5 more severe. For L_2, Block 5 experiences the most severe collapse, followed by Block 4. Block 1 L_2 remains consistently lowest, possibly due to VGG initialization.

F.3 CIFAR-100

CIFAR-100 experiments replicate the dual collapse and criterion-dependent threshold on a harder 100-class classification task, across both ResNet-18 and ResNet-34.

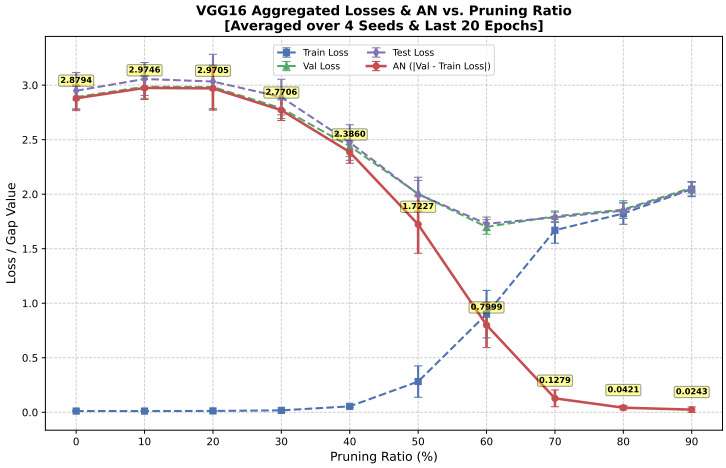


Figure 18: **Generalization metrics of VGG-16 across 4 random seeds.** AN index (red), test loss (purple), and validation loss (green) are shown alongside training loss (blue). Prior to 60% pruning, as memorization capacity of the deeper blocks is progressively stripped, test loss, validation loss, and AN index decline continuously. The test loss reaches its minimum at 60%, indicating optimal generalization. Beyond this threshold, test loss rebounds while AN continues to decrease—a consequence of impaired feature extraction rather than improved generalization.

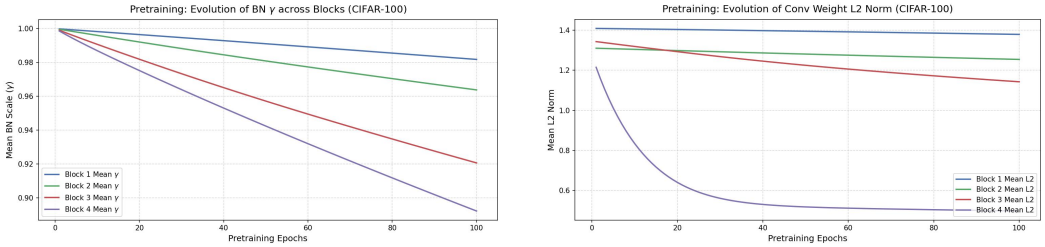
Table 7: **Architectural decomposition of VGG-16 internal prunable channel capacities.** To ensure dimensional alignment between Block 5 and the fully connected (FC) layers, we did not prune the final layer of Block 5. Consequently, the effective total number of prunable channels in Block 5 is 1024. Block 4 alone holds 41.38% of the total prunable channels in the network, while Block 5 holds 27.58%.

Block Name	**Width** (C)	**Prunable Conv Layers**	**Total Channels**	**Prunable Channels**	**% of Total**
Block 1	64	2	$64 \times 2 = 128$	128	3.45%
Block 2	128	2	$128 \times 2 = 256$	256	6.90%
Block 3	256	3	$256 \times 3 = 768$	768	20.69%
Block 4	512	3	$512 \times 3 = 1536$	1536	**41.38%**
Block 5	512	3-1	$512 \times 2 = 1024$	1024	27.58%
Total	**—**	**12**	**—**	**3712**	**100.00%**

* VGG-16 Block 5 contains 3 convolution layers, but the final layer is excluded from the prunable scope.

Table 8: **Layer-wise pruning statistics (mean $\pm$ std) for ResNet-34 on CIFAR-100 across multiple seeds.** Total prunable channels = 3776. At least one channel is retained in each convolution layer. Channel selection from 10% pruning onwards targets exclusively Block 4, with Blocks 2 and 3 entering selection after crossing the threshold 40.68%, at which point Block 4 is entirely exhausted.

Prune Ratio	**Total Pruned**	**Block 1 (C=192)**	**Block 2 (C=512)**	**Block 3 (C=1536)**	**Block 4 (C=1536)**
0.10	377 / 3776	0	0	0	377.0 ± 0.0 ($24.54\% \pm 0.00\%$)
0.20	755 / 3776	0	0	0	755.0 ± 0.0 ($49.15\% \pm 0.00\%$)
0.30	1132 / 3776	0	0	0	1130.0 ± 0.0 ($73.57\% \pm 0.00\%$)
0.40	1510 / 3776	0	0	0	1508.0 ± 0.0 ($98.18\% \pm 0.00\%$)
0.50	1888 / 3776	0	87.8 ± 2.9 ($17.14\% \pm 0.58\%$)	264.2 ± 2.9 ($17.20\% \pm 0.19\%$)	1533.0 ± 0.0 ($99.80\% \pm 0.00\%$)
0.60	2265 / 3776	0	127.0 ± 0.0 ($24.80\% \pm 0.00\%$)	600.0 ± 0.0 ($39.06\% \pm 0.00\%$)	1533.0 ± 0.0 ($99.80\% \pm 0.00\%$)
0.70	2643 / 3776	0	127.0 ± 0.0 ($24.80\% \pm 0.00\%$)	978.0 ± 0.0 ($63.67\% \pm 0.00\%$)	1533.0 ± 0.0 ($99.80\% \pm 0.00\%$)
0.80	3020 / 3776	0	127.0 ± 0.0 ($24.80\% \pm 0.00\%$)	1354.5 ± 0.5 ($88.18\% \pm 0.03\%$)	1533.0 ± 0.0 ($99.80\% \pm 0.00\%$)
0.90	3398 / 3776	31.0 ± 1.2 ($16.15\% \pm 0.64\%$)	294.0 ± 1.2 ($57.42\% \pm 0.24\%$)	1530.0 ± 0.0 ($99.61\% \pm 0.00\%$)	1533.0 ± 0.0 ($99.80\% \pm 0.00\%$)



(a) γ (b) L_2

Figure 19: **BN scale factor (γ) and L_2-norm trajectories for each ResNet-34 block during pre-train on CIFAR-100 dataset.** On ResNet-34, same with the situation on ResNet-18, dual collapse of γ and L_2-norm is replicated.

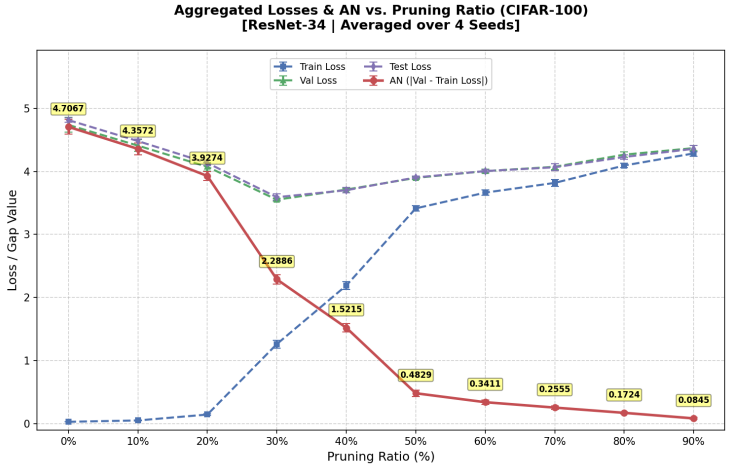


Figure 20: **Generalization metrics of ResNet-34/CIFAR-100 on L_2 criterion.** Model generalization improves as "noise sink" channels are progressively removed.

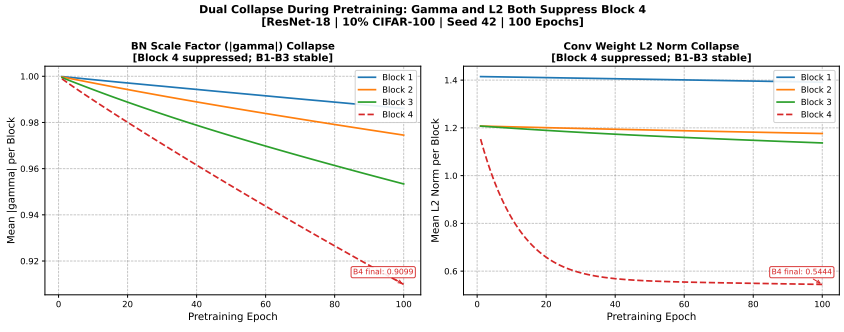


Figure 21: **BN scale factor (γ) (left) and L_2-norm (right) trajectories for each ResNet-18 block during pre-train on CIFAR-100 dataset.** On ResNet-18/CIFAR-100, dual collapse of γ and L_2-norm is replicated.

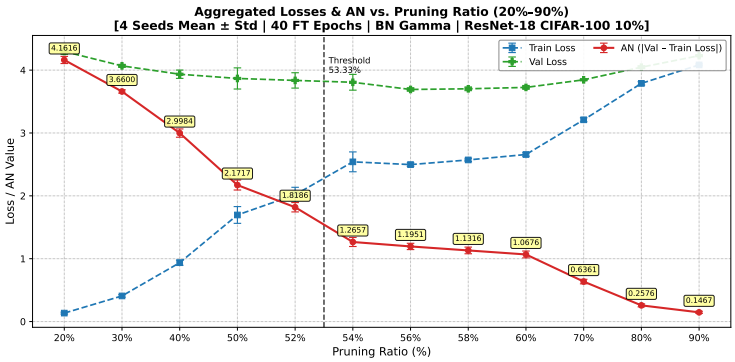


Figure 22: **Generalization metrics of ResNet-18/CIFAR-100 on L_2 criterion.** Model generalization improves as "noise sink" channels are progressively removed.

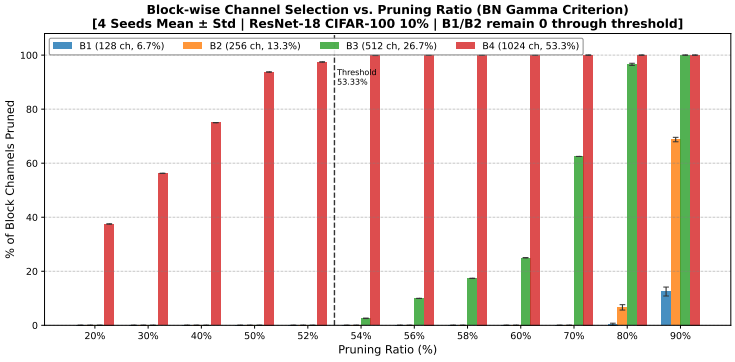


Figure 23: **Channel selection on ResNet-18/CIFAR-100 based on L_2 criterion.** Channel selection from 20% pruning onwards targets exclusively Block 4, with Block 3 entering selection after crossing the threshold 53.33%, at which point Block 4 is entirely exhausted.

G SCALE CHECK: DUAL COLLAPSE ON TINYIMAGENET

ResNet-18 trained on a 10% sample of TinyImageNet (64×64, 200 classes) confirms that the dual collapse and Block 4 suppression extend beyond CIFAR-scale imagery to a harder, higher-resolution task.

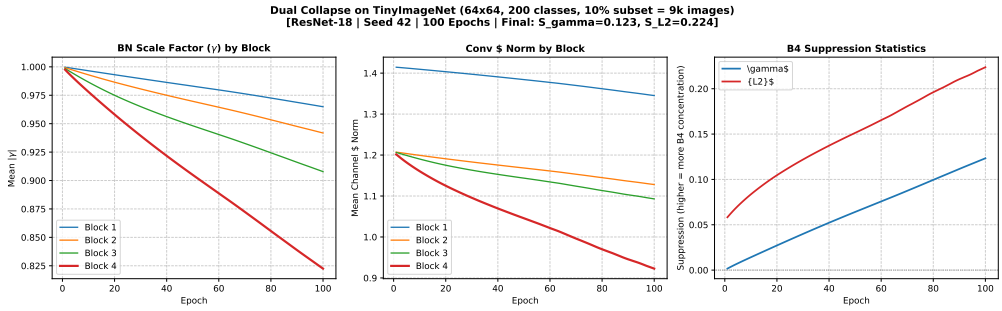


Figure 24: **Dual collapse on TinyImageNet confirms the structural finding at larger scale.** ResNet-18 trained on a 10% stratified sample of TinyImageNet (9,000 images, 64×64 resolution, 200 classes, seed 42, 100 pretrain epochs). Left and center: per-block BN γ and L_2-norm trajectories. Right: suppression statistics S_γ and S_{L2}. Block 4 exhibits a clear and progressive suppression below Blocks 1–3 on both metrics ($S_\gamma = 0.123$, $S_{L2} = 0.224$ at epoch 100), with the model reaching near-perfect training accuracy (99.9%) and 10.2% validation accuracy, confirming heavy overfitting. The depth-dependent structural differentiation observed on CIFAR-10/100 therefore extends to a higher-resolution, harder multi-class task.

H OPTIMIZER ABLATION: SGD VS ADAM

The figure below shows that L_2-based structural differentiation persists under SGD+momentum while γ collapse is Adam-specific, establishing that the weight-norm form of the signal is optimizer-independent.

I AN METRIC VALIDATION AGAINST DIRECT LOSS METRICS

The figure below confirms that the 53.33% threshold is visible in raw test and validation loss independently of AN, ruling out any measurement artifact introduced by the AN formulation.

J CONTROL EXPERIMENTS

Label-noise, data-fraction, and random-ranking controls establish that Block 4 suppression is a property of learned weight statistics, not an artifact of data quality, label corruption, or the pruning procedure itself.

K SECONDARY SHARP TRANSITION IN THE 20–30% RANGE

Under the L_2 and $L_2 \times \gamma$ criteria on ResNet-18 (Figures 2b, 2c), a secondary sharp transition is observed in the 20–30% range while only Block 4 channels are being removed. The mechanism is not yet fully understood and warrants further investigation. We propose that within Block 4, channels are not uniformly noisy. The lowest-ranked channels—those with the smallest magnitude scores—are the most specialized noise-fitters, with essentially no useful representational content. Removing them first eliminates the most extreme memorization units; when enough are removed, the remaining Block 4 channels must reorganize their shared load, which may create a brief instability that appears as a sharp event in the loss curve.

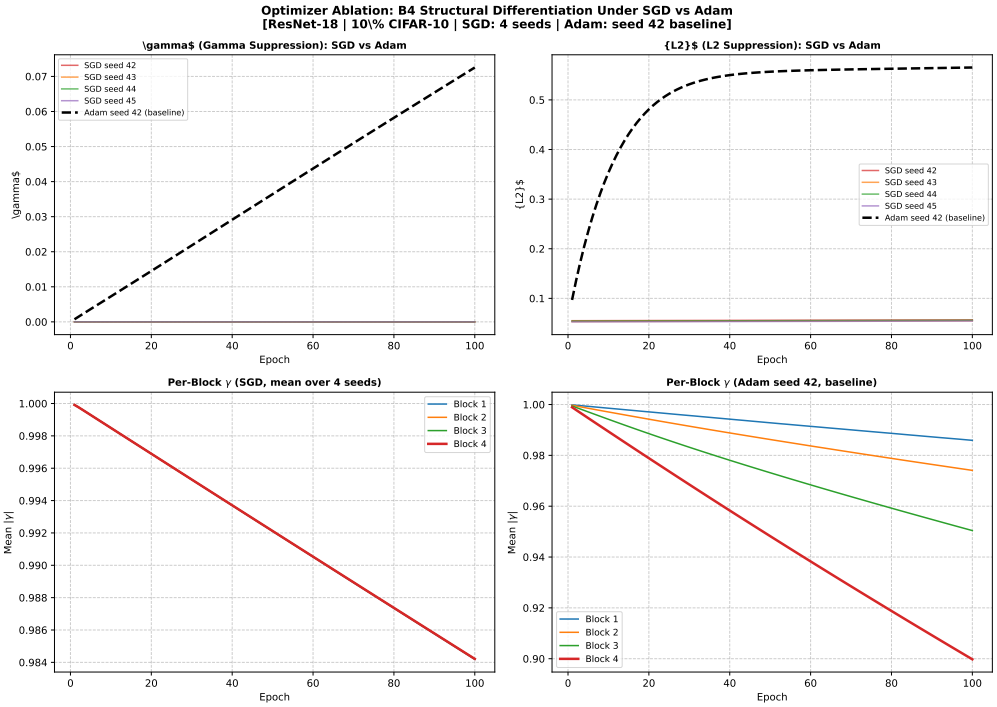


Figure 25: **Optimizer ablation: structural differentiation under SGD+momentum vs Adam.** ResNet-18 trained on 10% CIFAR-10. Top row: suppression statistics S_γ and S_{L2} comparing SGD (4 seeds, colored) against Adam baseline (seed 42, dashed black). Bottom row: per-block γ trajectories for SGD (mean over 4 seeds) and Adam (seed 42). Under Adam, both S_γ and S_{L2} are strongly positive (0.07 and 0.57 respectively at epoch 100). Under SGD, $S_\gamma \approx 0$ (no depth-dependent gamma differentiation) while $S_{L2} \approx 0.055$ (modest but consistent L2 suppression persists). This indicates that the L_2-based structural differentiation is optimizer-independent, while the gamma collapse is amplified by Adam's adaptive per-parameter learning rates. The L_2 criterion is therefore the more general signal.

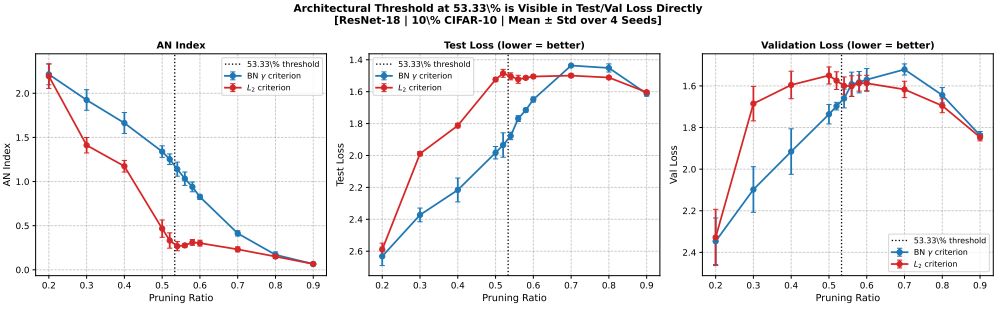


Figure 26: **The 53.33% architectural threshold is visible in test and validation loss directly, independently of the AN metric.** Each panel shows mean ± std across 4 seeds (ResNet-18, 10% CIFAR-10). Left: AN index (the metric used throughout the paper). Center: test loss (inverted; lower is better). Right: validation loss (inverted). All three metrics exhibit a consistent non-monotonic pattern with a regime change near 53.33%, confirming that the AN index is not a prerequisite for observing the threshold and does not introduce measurement artifacts.

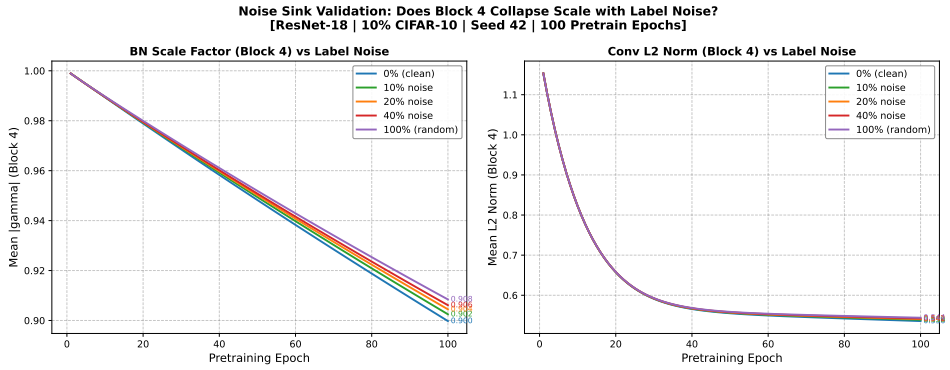


Figure 27: **Label-noise control: B4 dynamics do not scale with label corruption.** ResNet-18 trained on 10% CIFAR-10 with five label-noise levels (0%–100%). Left: mean $|\gamma_{B4}|$; right: mean L_2 norm of Block 4 channels. Neither metric decreases monotonically with noise level; the total variation from clean to fully randomised labels is less than 1%. This rules out label noise as a primary driver of the observed Block 4 suppression.

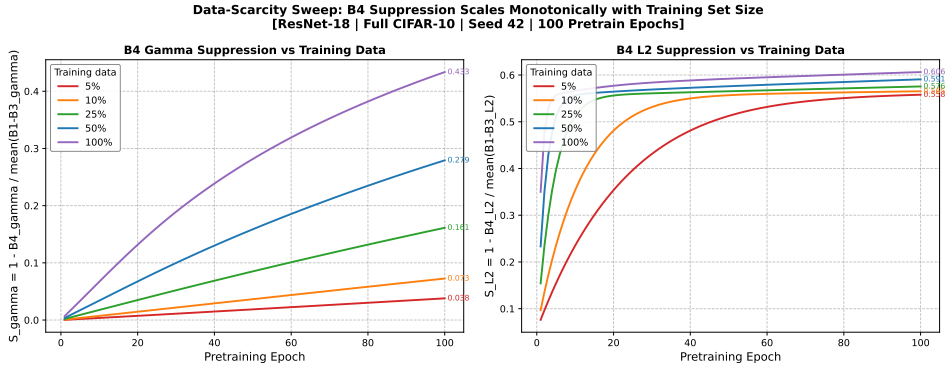


Figure 28: **Data-fraction sweep: B4 suppression strengthens with training data.** ResNet-18 trained on 5%, 10%, 25%, 50%, and 100% of full CIFAR-10 (seed 42, 100 pretrain epochs). The suppression statistic $S_\gamma = 1 - \gamma_{B4}/\text{mean}(\gamma_{B1:B3})$ increases monotonically from 0.038 at 5% to 0.434 at 100%. This shows that Block 4 suppression is a property of network convergence: as the model processes more training signal it develops stronger depth-dependent structural differentiation.

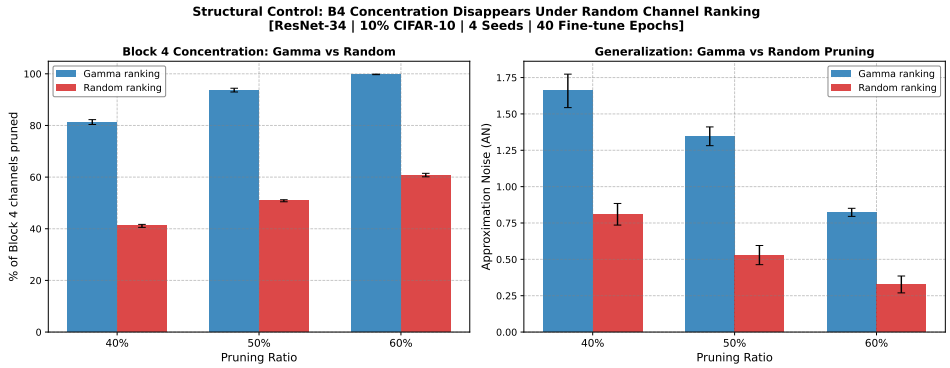


Figure 29: **Structural control: B4 concentration requires learned structure.** ResNet-34 pruned using (a) BN γ ranking and (b) uniform-random channel ranking, at 40%, 50%, and 60% ratios (4 seeds). Left: fraction of Block 4 channels removed. Right: AN index. Under random ranking, channel removal is approximately uniform across all blocks ($\approx$50% per block), whereas under γ ranking, Block 4 bears 81–100% of the removal load. This confirms that the preferential deepest-stage concentration is a property of the learned weight statistics, not of the pruning procedure itself.