Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

FILTER SPECTRAL EFFICIENCY (FSE): STRUCTURED NETWORK PRUNING *VIA* INTER-LAYER SPECTRAL COUPLING

Soumia Laroui[1], Hamida Seba[2], Karima Amrouche[1] and Yassine Ait Ali Yahia[1]

(Received: 12-Jul.-2026, Revised: 2-Sep.-2026 and 4-Sep.-2026, Accepted: 4-Sep.-2026)

ABSTRACT

Structured filter pruning is an effective approach for reducing the computational cost of deep neural networks while preserving efficient deployment on standard hardware. However, most existing pruning methods evaluate filters independently, overlooking the inter-layer spectral coupling between consecutive layers, while many high-performing approaches rely on activation statistics or calibration data. We propose Filter Spectral Efficiency (FSE), a data-free structured pruning framework that estimates filter importance through the spectral coupling between adjacent layers. FSE constructs singular-value-based filter descriptors, builds inter-layer coupling matrices, derives a spectral pruning criterion and allocates sparsity through a global adaptive budget. Extensive experiments on five CNN architectures for CIFAR-10 and ResNet-50 on ImageNet demonstrate that FSE consistently achieves a favorable accuracy-compression trade-off, outperforming representative structured pruning methods at comparable parameter and FLOP reductions. On CIFAR-10, FSE reduces the FLOPs of ResNet-56 by 59.8% with only a 0.07 percentage point accuracy drop and prunes GoogLeNet by 70.7% FLOPs while improving Top-1 accuracy over H Rank by 0.73 percentage points. On ImageNet, FSE improves the Top-1 accuracy of ResNet-50 by 1.14 percentage points over H Rank at a comparable FLOP reduction. These results demonstrate that inter-layer spectral coupling provides an effective and practical foundation for data-free structured network pruning.

KEYWORDS

Structured pruning, Data-free pruning, Inter-layer spectral coupling, Singular value decomposition, Deep neural networks.

1. INTRODUCTION

Deep neural networks (DNNs) have achieved remarkable success in visual-recognition tasks [1]-[2], but their high computational and memory demands hinder deployment on resource-constrained platforms such as mobile and embedded devices. Modern convolutional architectures, including ResNet [3], DenseNet [4] and GoogLeNet [5], contain millions of parameters and require billions of floating-point operations (FLOPs) per inference. Consequently, reducing computational cost while preserving predictive performance has become a central challenge in deep learning [6]. Beyond deployment efficiency, the increasing computational demands of deep learning have raised concerns about energy consumption and environmental impact. Green AI [7] advocates evaluating computational efficiency alongside accuracy, making model compression important for both resource-constrained deployment and sustainable AI.

Among model compression techniques [8]-[11], structured filter pruning has emerged as one of the most practical and hardware-friendly approaches, producing compact dense networks with genuine reductions in parameters, FLOPs, memory footprint and inference latency while remaining directly deployable on standard hardware [6], unlike unstructured pruning, which typically requires specialized sparse software or hardware to realize comparable acceleration [12]. This naturally raises the central question: Which filters can be removed with the least impact on the network's predictive performance?

Weight-based and data-dependent pruning criteria have been pursued in parallel rather than in succession. Weight-based methods estimate filter importance directly from the learned weights [13]-[14], whereas data-dependent methods exploit activation responses or calibration data [15]-[16]; both lines have continued to develop in parallel in subsequent work [17]-[19]. The data-free setting is nevertheless of practical importance, since data-dependent scoring adds computational cost and

1. S. Laroui, K. Amrouche and Y. Ait Ali Yahia are with Laboratoire de la Communication dans les Systèmes Informatiques (LCSI), Ecole nationale Supérieure d'Informatique (ESI), BP, 68M Oued-Smar, 16270 Alger, Algeria. Emails: {s_laroui; k_amrouche; y_ait_ali_yahia}@esi.dz
2. H. Seba is with Universite Claude Bernard Lyon 1, CNRS, INSA Lyon, LIRIS, Villeurbanne, France. Email: hamida.seba@univ-lyon1.fr

representative data may be unavailable for privacy, security or proprietary reasons. Independently of this distinction, many existing criteria assess filters primarily through individual properties or relationships within the same layer, without explicitly characterizing how the output of one layer couples to the representation consumed by the following layer. This leads to the following question: Can filter importance be estimated entirely from the network weights while explicitly modeling inter-layer spectral coupling between consecutive layers?

We hypothesize that filter importance is influenced not only by properties of individual filters and within layer redundancy, but also by the spectral coupling between a filter's output representation and those of the following layer. In other words, a filter may matter not merely because of its own magnitude or within-layer redundancy, but also because of how effectively its output interacts with the following layer. This hypothesis is tested in Sub-section 4.3: at the same FLOP reduction and using the same singular-value descriptors, replacing the coupling-based cost with a within-layer spectral score reduces Top-1 accuracy by 0.36 percentage points on ResNet-56/CIFAR-10, providing empirical support for the relevance of inter-layer spectral coupling to filter importance.

Motivated by this hypothesis, we introduce Filter Spectral Efficiency (FSE), a structured pruning framework with entirely data-independent filter-importance estimation. FSE first derives SVD-based spectral descriptors directly from the learned weights, then constructs an inter-layer coupling matrix the singular spectrum of which characterizes the spectral coupling between consecutive layers. From this spectrum, FSE estimates a per-filter spectral cost and integrates these costs into a global adaptive pruning budget that ranks filters across the network according to redundancy rather than fixed per-layer pruning ratios. Consequently, FSE prunes aggressively in redundant regions while preserving spectrally critical ones. All pruning decisions are made before any data is seen; only the subsequent fine-tuning uses training data, following standard structured pruning practice.

We evaluate FSE on CIFAR-10 [20] using representative CNN architectures spanning the dominant design paradigms: VGG-16 [21], ResNet-56/110 [3], DenseNet-40 [4] and GoogLeNet [5]. We further validate its scalability on the ImageNet ILSVRC-2012 benchmark [22] with ResNet-50. Across all settings, FSE achieves a favorable accuracy-compression trade-off, matching or outperforming recent structured-pruning methods, such as HRank [15], CHIP [23] and the Shapley-value criterion [17], while delivering genuine reductions in computational cost.

The main contributions of this paper are:

- A relational spectral filter-importance criterion. We propose an inter-layer spectral coupling criterion that estimates filter importance directly from the learned weights by modeling information transfer between consecutive layers.
- A global adaptive pruning strategy. FSE ranks filters network-wide according to spectral cost and allocates pruning adaptively under a single sparsity target, avoiding fixed layer-wise pruning ratios.
- A data-independent (data-free) pruning framework. All pruning decisions are derived directly from the learned weights, requiring no calibration data, activation statistics or feature-map collection.
- Comprehensive validation. Experiments on five CNN architectures for CIFAR-10 and ResNet-50 on ImageNet demonstrate that FSE matches or surpasses strong structured-pruning baselines while delivering genuine reductions in parameters and FLOPs.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed Filter Spectral Efficiency (FSE) framework, including the inter-layer coupling formulation, spectral cost, global adaptive pruning budget and the structural surgery stage for dependency-aware network reconstruction. Section 4 describes the experimental setup and presents comparisons with state-of-the-art methods together with ablation studies. Finally, Section 5 concludes the paper and discusses future-research directions.

2. RELATED WORK

Structured pruning removes coherent computational units, such as filters, channels or blocks, yielding smaller dense networks that accelerate directly on standard hardware, unlike unstructured pruning [12], the irregular sparsity of which often requires dedicated sparse kernels [6], [13]. FSE belongs to this family and evaluates compression through the actual reduction in FLOPs and parameters. We position

FSE along two axes: how filter importance is estimated across layers and whether the pruning decision requires data.

2.1 Modeling Cross-layer Interactions in Filter Pruning

Most existing criteria assess filters from individual properties or relationships within a layer. Magnitude-based methods [13] use ℓ_1/ℓ_2 norms, while FPGM [14] identifies within-layer geometric redundancy. Data-dependent methods, such as H Rank [15], CHIP [23] and FilterSketch [24], exploit activation or reconstruction information collected one layer at a time [25]-[26]. Learning-based approaches, such as GAL [16], DECORE [27], Auto- ℓ_0 pruning [18] and related methods [19], [28], optimize filter or layer gates, while explainability-based methods, such as Shapley-value pruning [17], estimate marginal contributions through attribution or sampling. These approaches generally do not explicitly model dependencies between consecutive layers. CIE [29] instead estimates cross-layer importance empirically through per-layer pruning and fine-tuning trials and uses the resulting contribution information to determine layer-wise sparsity. Thus, CIE incorporates cross-layer information through empirical pruning responses rather than through an explicit spectral coupling derived from the learned weights. FSE instead defines filter importance through the inter-layer spectral coupling between consecutive convolutional representations, computed analytically from the learned weights without activation sampling, attribution estimation or pruning trials.

2.2 Data-free Pruning

Beyond the cross-layer distinction, pruning methods also differ in whether their pruning decisions require data. Data-dependent criteria, such as HRank [15], CHIP [23] and Shapley-value pruning [17], require representative data to collect activations or evaluate filter contributions, introducing computational overhead and potential privacy or data-availability constraints. Data-free methods instead derive importance directly from the learned weights. Representative examples include magnitude pruning [13], FPGM [14] and Soft Filter Pruning [30]. CORING [31] applies higher-order SVD to measure within-layer filter similarity, while RASP [32] derives amplitude-based saliency from learned parameters and uses layer-sensitive regularization for global filter selection. However, removing data dependence from the importance criterion does not necessarily determine how sparsity should be allocated across layers. Fixed ratios, sensitivity analysis or additional optimization/search may still be required, as in Network Slimming [33], ABCPruner [34], AFPruner [35] and Auto- ℓ_0 Pruner [18]. FSE is data-independent throughout the pruning-decision stage: it derives filter importance from SVD-based descriptors of the learned weights and their inter-layer coupling and allocates the global sparsity budget directly from the resulting spectral cost, with only lightweight per-layer safety constraints. No calibration data, activation statistics, forward passes or auxiliary search/optimization are required to make the pruning decision; training data is used only in the subsequent fine-tuning stage.

2.3 Positioning of This Work

Despite their methodological diversity, the reviewed criteria differ in how they incorporate information beyond an individual filter. Most methods rely on within-layer properties or pairwise relationships, while a smaller body of work incorporates cross-layer information through empirical pruning responses. FSE departs from these approaches by using the spectral relationship between consecutive convolutional representations as the filter-importance criterion, while remaining fully data-independent at the pruning-decision stage. It differs from existing structured-pruning methods in three fundamental aspects:

1) an inter-layer spectral criterion that models the coupling between consecutive convolutional representations, providing a different formulation from existing cross-layer approaches;
2) a fully data-independent (data-free) pruning-decision mechanism that derives the required spectral information solely from the learned weights, without calibration data or activation statistics; and
3) an integrated global pruning strategy that combines the proposed spectral cost with adaptive network-wide sparsity allocation, rather than treating filter importance and pruning-budget allocation as separate steps.

Hybrid compression pipelines combine structured pruning with complementary techniques. NORTON [36] applies tensor decomposition followed by pruning using a principal-angle-based sub-space criterion, whereas CoDeC [37] clusters filters for coupled tensor decomposition using an SVD-based

sub-space measure. Both approaches exploit singular sub-spaces, but differ from FSE in scope and objective: NORTON prunes decomposition factors after tensor decomposition, CoDeC clusters filters within layers, whereas FSE uses rank-one SVD descriptors to model inter-layer coupling before structural pruning. Applying FSE as a pre-processing stage could reduce both channel dimensions before factorization; moreover, the remaining singular values obtained during the SVD, although discarded by the current FSE descriptor, could guide subsequent rank selection through their spectral decay. STPP [38] combines structured pruning with knowledge distillation; for FSE, the original unpruned network is already available as a teacher, making post-pruning feature-level distillation a natural extension. These combinations are left for future work.



Figure 1. Overview of the FSE pruning pipeline.

The following section presents the proposed Filter Spectral Efficiency (FSE) framework and details the formulation of the inter-layer coupling criterion, the global adaptive pruning budget and the overall pruning algorithm.

3. METHODOLOGY

Most structured-pruning criteria score each filter in isolation; by its ℓ_1 magnitude, its batch-normalization scale or its first-order effect on the loss, capturing how large a filter is, but not how its role relates to the following layer. Yet the output channels of one convolution are exactly the input channels of the next, so a filter's practical redundancy depends as much on this inter-layer relationship as on its own magnitude. Filter Spectral Efficiency (FSE) exploits this observation: it represents each filter by a compact spectral descriptor, measures how the descriptors of two consecutive layers couple and prunes the filters that contribute least to that coupling. FSE is data-free; operating purely on the learned weights, it yields a genuinely structured model compatible with standard inference devices.

The pipeline comprises five stages:

1) a per-filter spectral descriptor (Sub-section 3.1);
2) an inter-layer coupling matrix and its spectrum (Sub-section 3.2);
3) a per-filter spectral pruning cost (Sub-section 3.3);
4) a global spectral budget (Sub-section 3.4); and
5) the structural surgery that removes the selected channels (Sub-section 3.5).

The procedure is summarized in Algorithm 1 and illustrated in Figures 1 and 2-4. Notation. We consider a convolutional network with L prunable convolutional layers indexed by $\in \{0, \dots, L-1\}$. The weight tensor of layer l is

$$\boldsymbol{W}^{(l)} \in \mathbb{R}^{C_{\text{out}}^{(l)} \times C_{\text{in}}^{(l)} \times k \times k}, \tag{1}$$

where $C_{\text{out}}^{(l)}$ is the number of filters, $C_{\text{in}}^{(l)}$ the number of input channels and k the kernel size; the i th filter is $\boldsymbol{W}_i^{(l)} \in \mathbb{R}^{C_{\text{in}}^{(l)} \times k \times k}$. Each convolution is followed by a batch-normalization (BN) layer the parameters of which $\boldsymbol{\theta}_{\text{BN}}^{(l)} = \left(\boldsymbol{\gamma}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\mu}^{(l)}, \boldsymbol{v}^{(l)}\right)$ are pruned jointly with the corresponding filters. Finally, we define the protected set of layers

$$\mathcal{S} = \mathcal{S}_{\text{skip}} \cup \{0, L-1\}, \tag{2}$$

where $\mathcal{S}_{\text{skip}} \subseteq \{0, \dots, L-1\}$ is the sub-set of layers automatically detected as skip-connection (shortcut) layers during the Conv-BN parsing stage of Algorithm 1 and the indices 0 and $L-1$ denote the first and the last convolution, respectively. By convention we set $\kappa_i^{(l)} = +\infty$ for every i and every $l \in \mathcal{S}$, so that the output channels of a protected layer are never removed by the budget allocation of Eqs. (13)-(15); this preserves tensor-shape compatibility across residual branches and at the network boundaries.

3.1 Spectral Filter Descriptor

Existing weight-based pruning criteria commonly represent a filter by a scalar magnitude or other aggregate statistic, which may overlook the dominant structure encoded across its input channels. Since FSE aims to compare filters across layers using their learned weights, we require a compact representation that preserves the most informative spectral structure of each filter. We therefore use the leading singular component of each filter as its spectral descriptor, providing a data-independent representation of its dominant input-channel structure. Each filter is summarized by a compact vector capturing its dominant behaviour across input channels. The spatial dimensions are unfolded into a matrix

$$\boldsymbol{F}_i^{(l)} = \text{reshape}\left(\boldsymbol{W}_i^{(l)}\right) \in \mathbb{R}^{C_{\text{in}}^{(l)} \times k^2}, \tag{3}$$

where each row is the spatial weighting pattern of one input channel. Its singular value decomposition (SVD)

$$\boldsymbol{F}_i^{(l)} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}, \tag{4}$$

has singular values $s_1 \geq s_2 \geq \cdots$ with left singular vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$. The leading singular triplet $(s_1, \boldsymbol{u}_1, \boldsymbol{v}_1)$ provides the best rank-one approximation of the filter, so we retain only this dominant component:

$$\boldsymbol{c}_i^{(l)} = s_1 \boldsymbol{u}_1 \in \mathbb{R}^{C_{\text{in}}^{(l)}}. \tag{5}$$

Scaling the dominant direction $\boldsymbol{u}_1$ by s_1 preserves both the principal input-channel structure and its strength. Since the SVD fixes $\boldsymbol{u}_1$ only up to sign, we enforce the largest-magnitude entry to be positive,

$$\boldsymbol{c}_i^{(l)} \leftarrow \text{sgn}\left(c_{i,j^\star}^{(l)}\right) \boldsymbol{c}_i^{(l)}, j^\star = \arg\max_j \left|c_{i,j}^{(l)}\right|, \tag{6}$$

ensuring a deterministic representation for subsequent comparisons. This convention is a matter of bookkeeping rather than of geometry: the sign returned by an SVD routine is implementation-dependent, whereas the pruning cost of Eq. (12) is invariant to any per-filter sign choice (Remark 1). Hence, no preferred orientation is imposed on the dominant singular direction and the resulting pruning criterion is independent of the arbitrary sign convention.

$$\boldsymbol{D}^{(l)} = \left[\boldsymbol{c}_0^{(l)}, \ldots, \boldsymbol{c}_{C_{\text{out}}^{(l)}-1}^{(l)}\right]^{\top} \in \mathbb{R}^{C_{\text{out}}^{(l)} \times C_{\text{in}}^{(l)}}. \tag{7}$$

Each row of $\boldsymbol{D}^{(l)}$ acts as a fingerprint of a filter; filters with similar fingerprints capture similar dominant structures and may indicate functional redundancy (Figure 2).

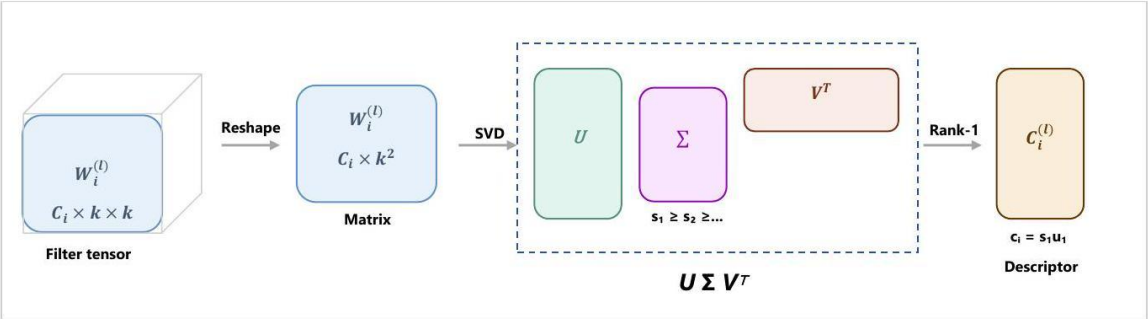


Figure 2. Construction of the per-filter spectral descriptor.

3.2 Inter-layer Coupling Matrix and Spectrum

A filter descriptor alone captures only the intrinsic structure of an individual filter and therefore cannot reveal how that filter interacts with the representation processed by the following layer. To explicitly capture this dependency while remaining data-independent, FSE compares the spectral descriptors of consecutive layers in a common representation space. We therefore construct an inter-layer coupling matrix based on cosine similarity and analyze its singular spectrum to characterize the dominant patterns of cross-layer alignment. First, when the descriptor dimensionality is large, descriptors are projected onto their leading principal directions to discard low-variance (noise-prone) components and reduce cost. Let $\boldsymbol{D}^{(l)\top} = \boldsymbol{U}_D \boldsymbol{\Sigma}_D \boldsymbol{V}_D^{\top}$ and let $d_{\max} = 16$ be the maximum retained dimension. The reduced

descriptor matrix is

$$\widetilde{\boldsymbol{D}}^{(l)} = \begin{cases} \boldsymbol{D}^{(l)}\boldsymbol{U}_D[:,1:d_{\max}], & C_{\text{in}}^{(l)} > d_{\max}, \\ \boldsymbol{D}^{(l)}, & \text{otherwise}. \end{cases} \tag{8}$$

Second, descriptors of adjacent layers may differ in dimensionality. Rather than truncating, the lower-dimensional descriptors are zero-padded to $d^\star = \max(d_1, d_2)$; since cosine similarity depends only on orientation, appending zeros leaves all pairwise similarities unchanged:

$$\widetilde{D} \mapsto \left[\widetilde{D} \| \mathbf{0}_{C_{\text{out}} \times (d^\star - d)}\right]. \tag{9}$$

Coupling is then measured through pairwise cosine similarities, appropriate here, because descriptors are compared by dominant direction rather than magnitude. Let $\widehat{\boldsymbol{D}}^{(l)}$ denote the ℓ_2-normalized rows of the padded $\widetilde{\boldsymbol{D}}^{(l)}, \widehat{\boldsymbol{d}}_\iota = \boldsymbol{d}_i / \|\boldsymbol{d}_i\|_2$. The coupling matrix is

$$\boldsymbol{M}^{(l)} = \widehat{\boldsymbol{D}}^{(l)}\left(\widehat{\boldsymbol{D}}^{(l+1)}\right)^\top \in \mathbb{R}^{C_{\text{out}}^{(l)} \times C_{\text{out}}^{(l+1)}}, M_{ij}^{(l)} = \cos\left(\boldsymbol{d}_i^{(l)}, \boldsymbol{d}_j^{(l+1)}\right), \tag{10}$$

where large values indicate that filter i in layer l and filter j in layer $l+1$ are strongly aligned. For a sequential pair of convolutional representations, Eq. (10) defines the pairwise spectral coupling used for filter-importance estimation. It is not intended to encode the complete computational dependency graph of the network. For architectures with non-sequential connectivity, such as DenseNet-40 and GoogLeNet, the FSE criterion retains this pairwise spectral formulation, while the subsequent structural-pruning stage uses DepGraph [39] to trace dependencies induced by dense concatenation, parallel branches and downstream operations and to propagate the selected channel-removal decisions accordingly.

The global structure of this alignment is revealed by the SVD

$$\boldsymbol{M}^{(l)} = \boldsymbol{U}^{(l)}\boldsymbol{\Sigma}^{(l)}\boldsymbol{V}^{(l)\top}, \sigma_1^{(l)} \geq \cdots \geq \sigma_r^{(l)} \geq 0, \tag{11}$$

the singular values $\left\{\sigma_k^{(l)}\right\}$ of which form the coupling spectrum. Each $\sigma_k^{(l)}$ measures the strength of an independent-alignment pattern, so the spectrum characterises the effective dimensionality of the inter-layer coupling. When spectral energy concentrates in few singular values, the coupling is effectively low-rank: many filters share the same dominant interaction patterns, indicating substantial redundancy.

3.3 Per-filter Pruning Cost

The coupling spectrum describes the importance of the different inter-layer alignment modes, but pruning requires converting this global spectral information into a filter-level quantity. A direct filter-level cost is therefore needed to rank filters according to their contribution to the cross-layer interaction. FSE obtains this cost from the participation of each filter in the singular modes, weighted by the corresponding singular values. A filter that contributes little to the spectrum carries little of the information exchanged between the two layers and is a pruning candidate, whereas a strongly contributing filter is structurally important. Reusing the SVD of (11), we define the pruning cost of filter i as z

$$\kappa_i^{(l)} = \sum_k \left(U_{ik}^{(l)}\right)^2 \sigma_k^{(l)}. \tag{12}$$

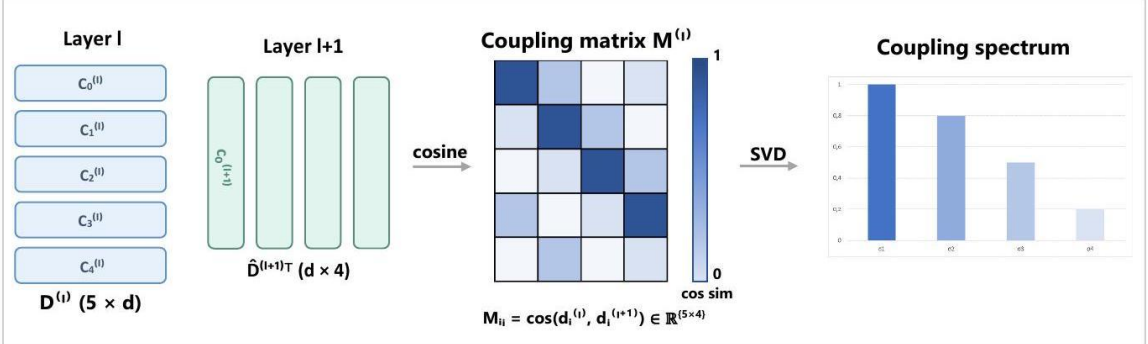


Figure 3. A representative inter-layer coupling matrix and its spectrum.

Here, $\left(U_{ik}^{(l)}\right)^2$ measures the participation of filter i in spectral mode k and $\sigma_k^{(l)}$ weights that mode by its significance, so filters participating strongly in dominant modes receive a large cost and are preferentially retained. The cost requires only $\mathcal{O}(C)$ per-layer computation, yet Proposition 1 shows it exactly decomposes the nuclear norm of the coupling matrix across filters, providing a principled rather than heuristic justification.

Protected filters. For any protected layer $l \in \mathcal{S}$ and for the final layer where no forward coupling is defined, all filters receive infinite cost, $\kappa_i^{(l)} = +\infty$, guaranteeing that they are never removed while keeping a uniform framework across layers.

3.4 Global Spectral Budget

A filter-level importance score does not by itself determine how many filters should be removed from each layer. Fixed layer-wise ratios impose the same pruning pressure on layers with potentially different levels of spectral redundancy, while separate sensitivity or search procedures introduce additional computational cost. Since FSE already provides a spectral cost for every filter, we use these costs to allocate a single global pruning budget across the network, allowing the pruning profile to emerge from the measured spectral redundancy. Let $\mathcal{E} = \left\{(l,i): \kappa_i^{(l)} < \infty\right\}$ be the eligible filters and $s \in [0,1]$ the target sparsity. The number of filters to remove is

$$T = \lfloor s|\mathcal{E}| \rfloor. \tag{13}$$

To prevent collapsing a layer, each layer l (with n_l eligible filters) is capped:

$$\text{cap}_l = \begin{cases} 0, & n_l \le 1, \\ \max\left(0, \min\left(n_l - 1, \lfloor n_l \rho_{\max} \rfloor\right)\right), & n_l > 1, \end{cases} \tag{14}$$

where $\rho_{\max}$ (default 0.8) is the maximum removable fraction per layer; this keeps at least one filter per non-trivial layer and avoids over-concentrated pruning. Eligible filters are ranked by ascending cost and selected greedily until the budget is exhausted, subject to the caps:

$$\mathcal{P} = \left\{(l,i): (l,i) \text{ visited in ascending order of } \kappa_i^{(l)}, |\mathcal{P}| \le T, |\{(l,\cdot) \in \mathcal{P}\}| \le \text{cap}_l\right\}. \tag{15}$$

The effective sparsity is $|\mathcal{P}|/|\mathcal{E}|$; if the caps prevent the target from being reached, the shortfall is reported so the user may relax $\rho_{\max}$ or lower s. Because ranking is network-wide, FSE removes more filters from spectrally redundant layers and fewer from important ones, so the allocation emerges from the spectral costs rather than manual ratios (Figure 4). The greedy rule is also optimal: since costs are additive and feasibility is encoded by the caps, selecting the lowest-cost feasible filters minimizes the total pruning cost over all feasible selections of equal cardinality (Proposition 2).

3.5 Structural Surgery

Selecting filters is not sufficient to obtain a valid compact network, because removing an output channel from one layer can change the input-channel structure of downstream layers. This issue is particularly important for architectures with complex connectivity, such as dense concatenations and parallel branches. FSE therefore performs dependency-aware structural surgery to preserve the computational graph after pruning; for DenseNet-40 and GoogLeNet, DepGraph [39] is used to trace the resulting channel dependencies and propagate the pruning decisions to affected layers.

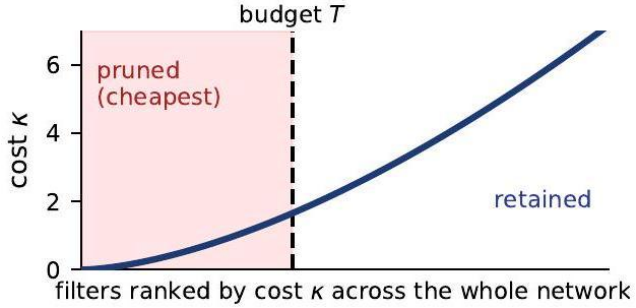


Figure 4. Global cost-ranked filter budget.

Algorithm 1. Filter Spectral Efficiency (FSE) pruning.

Require: network f; target sparsity s; cap $\rho_{\max}$
extract Conv-BN pairs and build the protected set $\mathcal{S}$
for each layer l **do**
  compute descriptors $\boldsymbol{D}^{(l)}$ via Eqs. (3)-(7)
end for
for $l = 0$ to $L - 2$ **do**
  form $\boldsymbol{M}^{(l)}$ (Eq. (10)) and its SVD (Eq. (11))
end for
for each layer l **do**
  if $l \in \mathcal{S}$ or coupling undefined **then**
    $\kappa_i^{(l)} \leftarrow \infty \forall i$
  else
    compute $\kappa^{(l)}$ via Eq. (12)
  end if
end for
select $\mathcal{P}$ by the global budget (Eqs. (13)-(15))
perform structural surgery (Eqs. (16)-(18))
return pruned network (After fine-tuned)

The final stage turns the pruning set $\mathcal{P}$ into a physically smaller network. For each layer l, let $\mathcal{P}_l = \{i: (l,i) \in \mathcal{P}\}$ be the pruned output channels and $\mathcal{K}_l = \left\{0, \dots, C_{\text{out}}^{(l)} - 1\right\} \setminus \mathcal{P}_l$ the retained ones. The pruned channels are removed from the kernel and BN parameters,

$$\boldsymbol{W}^{(l)} \leftarrow \boldsymbol{W}^{(l)}[\mathcal{K}_l, :, :, :], \tag{16}$$

$$\boldsymbol{\theta}_{\text{BN}}^{(l)} \leftarrow \boldsymbol{\theta}_{\text{BN}}^{(l)}[\mathcal{K}_l]. \tag{17}$$

and the decision is propagated to every downstream convolution g consuming layer l :

$$\boldsymbol{W}^{(g)} \leftarrow \boldsymbol{W}^{(g)}[:, \mathcal{K}_l, :, :]. \tag{18}$$

4. EXPERIMENTS

We evaluate FSE on CIFAR-10 and ImageNet using six CNN architectures, compare it with representative structured-pruning methods and conduct ablation studies on the proposed inter-layer spectral criterion and global spectral budget.

4.1 Experimental Setup

Datasets. Experiments are conducted on CIFAR-10 [20] and ImageNet ILSVRC-2012 [22]. CIFAR-10 consists of 60,000 color images of size 32 × 32 from 10 classes, split into 50,000 training and 10,000 test images. ImageNet ILSVRC-2012 contains approximately 1.28 million training images and 50,000 validation images across 1,000 classes with the shape of 224 × 224.

Architectures. On CIFAR-10, we evaluate FSE on ResNet-56, ResNet-110 [3], DenseNet-40 [4], GoogLeNet [5] and VGG-16 [21], covering residual, densely connected, inception-style and plain CNN architectures. For large-scale evaluation, we also prune ResNet-50 [3] on ImageNet.

Implementation details. FSE operates directly on trained weights without calibration data. On CIFAR-10, pruned models are fine-tuned for 160 epochs using SGD (momentum 0.9, weight decay 5×10^{-4}), an initial learning rate of 0.01 with cosine decay and a batch size of 128. On ImageNet, ResNet-50 is finetuned for 90 epochs using SGD with the same momentum and weight decay, an initial learning rate of 0.01 with cosine annealing, a batch size of 256 and AMP. Unless stated otherwise, FSE uses the proposed pruning criterion, a descriptor cap $d_{\max} = 16$ and a maximum per-layer pruning ratio $\rho_{\max} = 0.8$.

Metrics and protocol. On CIFAR-10, we report top-1 accuracy together with parameter and FLOP reductions relative to the unpruned baseline. On ImageNet, we report top-1, top-5 accuracy, parameter reduction and FLOP reduction. For baselines, we report results from the original papers when available. A dash "-" indicates that the corresponding metric was not reported. We ensure that comparisons use

the same network architectures.

Table 1. Comparison with state-of-the-art pruning methods on VGG-16/CIFAR-10.

Model	Method	Top-1 Acc. ↑ / ↓ (%)	Params (M)	Params ↓ (%)	FLOPs (M)	FLOPs ↓ (%)
VGG-16	Baseline	93.02 (0.00)	14.73	0.00	314.31	0.00
	L1 [13]	93.40 (+0.38↑)	5.40	64.00	206.00	34.20
	FPGM [14]	94.00 (+0.42↑)	-	-	206.00	34.20
	DECORE-500 [27]	94.02 (+0.06↑)	5.54	63.00	203.08	35.30
	FSE (Ours)	93.43 (+0.41↑)	8.14	44.72	200.88	36.09
	GAL-0.05 [16]	92.03 (-1.93↓)	3.36	77.60	189.49	39.60
	GAL-0.1 [16]	90.73 (-3.23↓)	2.67	82.20	171.89	45.20
	Network slimming [33]	93.80 (+0.14↑)	2.30	88.50	-	51.00
	FSE (Ours)	**93.20 (+0.18↑)**	**3.93**	**73.29**	**124.23**	**60.47**
	FICP [19]	93.72 (-0.24↓)	-	85.90	-	63.20
	FSE (Ours)	**93.13 (+0.11↑)**	**3.25**	**77.95**	**111.98**	**64.37**
	HRank [15]	91.23 (-2.73↓)	1.78	92.00	73.70	76.50
	Shapley [17]	93.14 (-0.82↓)	1.90	87.30	67.09	78.60
	FSE (Ours)	**92.15 (-0.87↓)**	**1.38**	**90.66**	**66.94**	**78.70**
	Auto ℓ_0 Pruner-0.5 [18]	90.55 (-2.47↓)	1.65	88.80	34.61	88.99
	FSE (Ours)	**90.65 (-2.37↓)**	**1.07**	**92.80**	**34.36**	**89.10**

Hardware and software setup. All experiments were conducted on a workstation equipped with a single NVIDIA GeForce RTX 5090 GPU with 32 GB of VRAM (CUDA 12.9), an AMD EPYC 9654 processor with 96 CPU cores, 76.5 GB of system memory and 350 GB of SSD storage. The experiments were performed under Ubuntu 24.04.4 LTS using Python 3.12.14 and PyTorch 2.11.0+cu128, with an NVIDIA driver version 610.57.04.

4.2 Comparison with State-of-the-Art

4.2.1 Results on CIFAR-10

VGG-16. As a highly over-parameterized architecture, VGG-16 is a standard benchmark for structured pruning. As summarized in Table 1, FSE achieves a favorable accuracy-efficiency trade-off across all compression regimes. At moderate compression, 36.09% FLOPs, it improves the baseline by +0.41%, matching FPGM's +0.42% while pruning more FLOPs than the ℓ_1-norm and DECORE-500. At intermediate ratios, it remains the only method staying above baseline while saving more compute: +0.18% at 60.47% FLOPs *vs.* Network Slimming's +0.14% at only 51% and 93.13% at 64.37% FLOPs *vs.* FICP's -0.24% at 63.20%. Under aggressive pruning, FSE stays competitive: at 78.70% FLOPs, it reaches 92.15%, -0.87%, exceeding H Rank by 1.68 pp and matching Shapley's -0.82% with fewer parameters, 1.38M *vs.* 1.90M, yet without any calibration data or activation statistics. At the extreme 89.10% FLOP setting, it still surpasses Auto- ℓ_0 Pruner-0.5 on all three metrics: 90.65% *vs.* 90.55% accuracy, 1.07M *vs.* 1.65M parameters. Notably, the FLOP reduction consistently exceeds the parameter reduction, indicating that FSE preferentially removes filters in computationally dominant layers, a behavior well suited to latency-bound edge deployment.

ResNet-56. ResNet-56 is a far more challenging benchmark than VGG-16 owing to its compact design and low redundancy. As reported in Table 2, FSE sustains a favorable accuracy-compression tradeoff across all ratios. At moderate compression, it reaches 93.27%, +0.01 pp, at 53.05% FLOPs, the largest reduction in its group. As the budget grows, it keeps the lead: -0.07% at 59.84% FLOPs, vs. ℓ_0 Pruner's -0.18% and 93.03% at 62.45% FLOPs, exceeding both Auto- ℓ_0 Pruner-0.3 92.96% and Direct-Indirect Filter Pruning 93.04% at higher compression. Under aggressive pruning, the gap widens: at ≈ 72% FLOPs, it attains 92.43%, beating Shapley 92.20% and CHIP 92.05%, with the largest parameter reduction, 72.47%; and at 74.48% FLOPs, it retains 92.07%, surpassing HRank by 1.35 pp with larger cuts in both parameters, 76.09% *vs.* 68.10% and FLOPs. These results demonstrate that FSE generalizes across architectures rather than exploiting architecture-specific redundancy.

Table 2. Comparison with state-of-the-art pruning methods on ResNet-56 and ResNet110/CIFAR-10.

Model	Method	Top-1 Acc. ↑/↓ (%)	Params (M)	Params ↓ (%)	FLOPs (M)	FLOPs ↓ (%)
ResNet-56	Baseline	93.26 (0.00)	0.85	0.00	127.62	0.00
	L1 [13]	93.06 (-0.20↓)	0.73	14.12	90.90	27.60
	GAL-0.6 [16]	92.98 (-0.28↓)	0.75	11.80	78.30	37.60
	SFP [30]	93.35 (-0.24↓)	-	-	59.40	52.60
	FPGM [14]	93.49 (-0.10↓)	-	-	59.40	52.60
	FSE (Ours)	**93.27 (+0.01↑)**	**0.52**	**39.50**	**59.92**	**53.05**
	ℓ_0 Pruner (63% pruned) [18]	93.08 (-0.18↓)	0.28	67.06	51.38	59.74
	FSE (Ours)	**93.19 (-0.07↓)**	**0.42**	**50.49**	**51.25**	**59.84**
	GAL-0.8 [16]	90.36 (-2.90↓)	0.29	65.90	49.99	60.20
	Direct-Indirect Filter Pruning [28]	93.04 (-0.60↓)	-	-	-	61.50
	Auto ℓ_0 Pruner-0.3 [18]	92.96 (-0.30↓)	0.31	63.53	48.47	62.02
	FSE (Ours)	**93.03 (-0.23↓)**	**0.38**	**55.15**	**47.92**	**62.45**
	CHIP [23]	92.05 (-1.21↓)	0.24	71.80	34.79	72.30
	Shapley [17]	92.20 (-1.06↓)	0.24	71.80	34.79	72.30
	FSE (Ours)	**92.43 (-0.83)**	**0.23**	**72.47**	**35.32**	**72.32**
	HRank [15]	90.72 (-2.54↓)	0.27	68.10	32.52	74.10
	FSE (Ours)	**92.07 (-1.19↓)**	**0.20**	**76.09**	**32.56**	**74.48**
ResNet-110	Baseline	93.50 (0.00)	1.73	0.00	257.09	0.00
	L1 [13]	93.30 (-0.20↓)	1.16	32.60	155.00	38.70
	SFP [30]	93.35 (+0.18↑)	-	-	152.20	40.80
	GAL-0.5 [16]	92.55 (-0.95↓)	0.95	44.80	130.20	48.50
	FPGM [14]	93.74 (+0.06↑)	-	-	121.00	52.30
	FSE (Ours)	**93.80 (+0.30↑)**	**1.04**	**39.66**	**120.28**	**53.21**
	FilterSketch [24]	93.44 (-0.06↓)	0.69	59.90	92.48	63.30
	FSE (Ours)	93.46 (-0.04↓)	0.70	59.74	88.35	65.63
	HRank [15]	92.65 (-0.85↓)	0.53	68.70	79.30	68.60
	FSE (Ours)	**93.31 (-0.19↓)**	**0.58**	**66.34**	**78.82**	**69.34**
	Shapley [17]	93.14 (-0.36↓)	0.54	68.60	72.83	71.30
	FSE (Ours)	**93.18 (-0.29↓)**	**0.52**	**69.76**	**73.13**	**71.56**
	DECORE-175 [27]	92.71 (-0.79↓)	0.35	79.60	58.37	76.90
	FSE (Ours)	**92.93 (-0.57↓)**	**0.47**	**72.86**	**58.18**	**77.37**

ResNet-110. The deeper ResNet-110 confirms these findings. At moderate compression, 53.21% FLOPs, FSE attains the highest accuracy of all methods, 93.80%, +0.30 pp, exceeding FPGM by 0.24 pp, SFP by 0.18 pp and the ℓ_1-norm criterion by 0.20 pp while pruning the most FLOPs. It stays ahead of recent feature-based methods: 93.46% at 65.63% FLOPs vs. FilterSketch, 93.44% at 63.30% FLOPs and +0.66 pp over HRank at 69.34% FLOPs. It again outperforms Shapley at $\approx$ 71.5% FLOPs in accuracy, 93.18% *vs.* 93.14%, as well as in parameters and FLOPs. Even at the most aggressive 77.37% FLOP reduction, it retains 92.93%, surpassing DECORE-175 by 0.22 pp. These results indicate that inter-layer spectral coupling provides a stable importance measure as residual networks deepen.

GoogLeNet. GoogLeNet is harder to prune than sequential or residual nets: its Inception modules concatenate multiple branches before downstream consumption, so pruning must preserve both interlayer flow and channel consistency (Table 3). At moderate compression, 66.72% FLOPs, FSE attains the highest accuracy of all methods, 94.91%, -0.14%, exceeding ABCPruner by 1.07 pp with more parameters removed, 63.91% *vs.* 60.14% and FICP by 0.16 pp at a comparable budget. The lead widens under aggressive pruning: 94.80% at 70.68% FLOPs, +0.73 pp over HRank and at the highest setting, 73.95% FLOPs, it matches Shapley within 0.03 pp, 94.66% *vs.* 94.69%, while producing a far more compact model, 1.78M *vs.* 2.10M params, 71.18% *vs.* 65.90% removed. Across all points, FSE stays within 0.39 pp of the 95.05% baseline while cutting nearly 74% of the compute, confirming that inter-layer spectral coupling remains effective even under complex multi-branch feature aggregation.

DenseNet-40. DenseNet-40 provides a particularly challenging benchmark for structured pruning,

because its dense feature reuse leaves limited redundancy. As shown in Table 4, FSE consistently achieves the best accuracy-compression trade-off across both evaluated operating points. At moderate compression, 41.01% FLOP reduction, FSE attains the highest Top-1 accuracy 94.30%, slightly outperforming HRank while operating at a marginally higher FLOP reduction. Under more aggressive pruning, 61.07% FLOP reduction, FSE further improves upon HRank and GAL-0.05, demonstrating that the proposed inter-layer spectral criterion remains effective even when redundancy is limited. Unlike VGG-16, FSE prioritizes computational efficiency over parameter reduction, achieving the largest FLOP reduction while maintaining the highest predictive performance. Overall, these results demonstrate that FSE generalizes effectively to densely connected architectures, providing the strongest overall accuracy-compression trade-off despite the limited redundancy of DenseNet-40.

Table 3. Comparison with state-of-the-art pruning methods on GoogLeNet/CIFAR-10.

Model	Method	Top-1 Acc. ↑/↓ (%)	Params (M)	Params ↓ (%)	FLOPs (B)	FLOPs ↓ (%)
GoogLeNet	Baseline	95.05 (0.00)	6.15	0.00	1.52	0.00
	L1 [13]	94.54 (-0.51↓)	3.51	42.90	1.02	32.90
	GAL-0.05 [16]	93.93 (-1.12↓)	3.12	49.30	0.94	38.20
	GAL-ApoZ [16]	92.11 (-2.94↓)	2.85	53.70	0.76	50.00
	HRank [15]	94.53 (-0.52↓)	2.74	55.40	0.69	54.90
	FICP [19]	94.75 (-0.30↓)	-	55.00	-	64.50
	ABCPruner [34]	93.84 (-0.21↓)	2.46	60.14	0.51	66.56
	FSE (Ours)	**94.91 (-0.14↓)**	**2.23**	**63.91**	**0.52**	**66.72**
	HRank [15]	94.07 (-0.98↓)	1.86	69.80	0.45	70.40
	FSE (Ours)	**94.80 (-0.25↓)**	**2.02**	**67.22**	**0.45**	**70.68**
	Shapley [17]	94.69 (-0.36↓)	2.10	65.90	0.40	73.90
	FSE (Ours)	**94.66 (-0.39↓)**	**1.78**	**71.18**	**0.40**	**73.95**

Table 4. Comparison with state-of-the-art pruning methods on DenseNet-40/CIFAR-10.

Model	Method	Top-1 Acc. ↑/↓ (%)	Params (M)	Params ↓ (%)	FLOPs (M)	FLOPs ↓ (%)
DenseNet-40	Baseline	94.81 (0.00)	1.06	0.00	292.50	0.00
	GAL-0.01 [16]	94.29 (-0.52↓)	0.67	35.60	182.92	35.30
	HRank [15]	94.24 (-0.57↓)	0.66	36.50	167.41	40.80
	FSE (Ours)	**94.30 (-0.51↓)**	**0.77**	**26.83**	**172.56**	**41.01**
	GAL-0.05 [16]	93.53 (-1.28↓)	0.45	56.70	128.11	54.70
	HRank [15]	93.68 (-1.13↓)	0.48	53.80	110.15	61.00
	FSE (Ours)	**93.79 (-1.02↓)**	**0.50**	**52.67**	**113.86**	**61.07**

4.2.2 Results on ImageNet

ResNet-50. We finally evaluate the scalability of FSE on the large-scale ImageNet benchmark using ResNet-50. For completeness, RASP [32] and CORING [31] are not included in Table 5, because their pruning protocols are not directly comparable with the single-shot pruning protocol used in this study: RASP relies on regularization-based model selection, whereas CORING employs an iterative K-shots procedure with intermediate fine-tuning. As shown in Table 5, FSE consistently achieves a favorable accuracy-compression trade-off across all compression regimes.

At moderate compression, 46.60% FLOP reduction, FSE attains the best overall trade-off, achieving 74.81% Top-1 and 92.29% Top-5 accuracy while providing the largest parameter and FLOP reductions among comparable methods. Compared with the recent Shapley-based approach, FSE delivers higher Top-1 and Top-5 accuracy together with greater compression, without requiring calibration data or activation statistics. As compression increases, FSE continues to outperform competing methods. At 57.53% FLOP reduction, it achieves the highest accuracy among methods with comparable compression, surpassing AFPruner, FPGM, ABCPruner and Auto- ℓ_0 Pruner. Under severe compression, 62.45% FLOP reduction, FSE further exceeds HRank by 1.14 percentage points in Top-1 accuracy while simultaneously achieving larger parameter and FLOP reductions. Overall, these results

demonstrate that the proposed inter-layer spectral criterion scales effectively from CIFAR-10 to ImageNet, consistently providing an excellent balance between predictive accuracy and computational efficiency through entirely data-free pruning.

Table 5. Comparison with state-of-the-art pruning methods on ILSVRC-2012 using the ResNet50 architecture.

Model	Method	Top-1 Acc. ↑/↓ (%)	Top-5 Acc. (%)	Params (M)	Params ↓ (%)	FLOPs (B)	FLOPs ↓ (%)
ResNet-50	Baseline	76.15 (0.00)	92.88	25.56	0.00	4.13	0.00
	SSS-32 [40]	74.18 (-1.94↓)	91.91 (-0.95↓)	18.60	27.06	2.82	31.05
	He et al. [26]	72.30 (-3.83↓)	90.80 (-2.06↓)	-	-	2.73	33.30
	ThiNet [25]	72.04 (-1.57↓)	90.67 (-0.69↓)	16.94	33.7	-	36.80
	SFP [30]	74.61 (-1.54↓)	92.06 (-0.81↓)	-	-	-	41.80
	DECORE-6 [27]	74.58 (-1.57↓)	92.18 (-0.69↓)	14.10	44.71	2.36	42.30
	GAL-0.5 [16]	71.95 (-4.18↓)	90.94 (-1.92↓)	21.20	16.86	2.33	43.03
	Shapley [17]	74.70 (-1.45↓)	92.18 (-0.69↓)	15.09	40.82	2.28	44.25
	FilterSketch-0.6 [24]	74.68 (-1.45↓)	92.17 (-0.69↓)	14.53	43.00	2.23	45.50
	FSE (Ours)	**74.81 (-1.34↓)**	**92.29 (-0.59↓)**	**14.03**	**45.10**	**2.21**	**46.60**
	AFPruner [35]	74.23 (-1.78↓)	91.89 (-1.07↓)	12.00	53.05	2.06	50.18
	FPGM [14]	74.13 (-2.02↓)	92.87 (-0.93↓)	-	-	1.92	53.50
	ABCPruner [34]	73.86 (-2.15↓)	91.69 (-1.27↓)	11.75	54.02	1.89	54.29
	Auto ℓ_0 Pruner-0.1 [18]	74.15 (-2.00↓)	91.70 (-1.18↓)	12.50	51.10	1.77	57.14
	FSE (Ours)	**74.43 (-1.77↓)**	**91.92 (-0.96↓)**	**12.92**	**49.43**	**1.76**	**57.53**
	HRank [15]	71.98 (-4.17↓)	91.01 (-1.86↓)	13.77	46.10	1.55	62.10
	FSE (Ours)	**73.12 (-3.03↓)**	**91.29 (-1.58↓)**	**12.16**	**52.41**	**1.55**	**62.45**

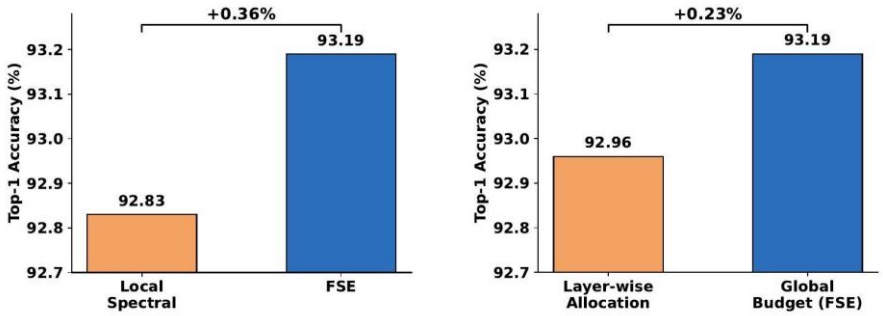
Compared with recent data-free pruning approaches, such as RASP [32] and CORING [31], FSE offers a different combination of structural information and pruning simplicity. RASP derives a data-independent amplitude-based saliency criterion, but its pruning framework involves additional regularization and model-selection steps followed by fine-tuning, whereas CORING uses higher-order SVD to identify redundant filters through within-layer similarity and adopts an iterative K-shot pruning strategy. In its main experiments, CORING uses $K = 15$, allowing fine-tuning between successive pruning shots, with a total of 100 calibration epochs allocated to the K-shot procedure [31]. In contrast, FSE determines the final pruning structure directly from the learned weights and performs fine-tuning only after the complete pruning decision. Therefore, the slightly higher accuracy achieved by RASP or CORING can be viewed in the context of their respective pruning protocols, which allow additional model adaptation during the pruning process. FSE instead targets a favorable accuracy-compression trade-off while maintaining a direct, lightweight and analytically defined pruning decision based solely on the learned weights.

4.3 Ablation Studies

We isolate the contribution of the two core components of FSE; the inter-layer spectral coupling cost and the global adaptive budget through controlled, single-variable ablations on ResNet-56/CIFAR-10. In each experiment, we change exactly one component and keep everything else identical: the SVD filter descriptors, the per-layer pruning caps, the skip-connection protection and the fine-tuning schedule. Crucially, every variant is evaluated at the same FLOP reduction (59.84%) rather than at the same nominal sparsity, because two criteria that prune the same number of filters generally remove different amounts of compute; matching FLOPs ensures that the accuracy gaps reported in Figs. 5a and 5b are not confounded by the operating point.

Effect of inter-layer spectral coupling. Replacing the coupling-based cost with a purely local perfilter spectral score, the descriptor norm $\|d_i\|_2$; i.e., each filter's own leading singular value, with the inter-layer coupling matrix discarded lowers Top-1 accuracy from 93.19% to 92.83% (-0.36 pp, Fig. 5a). Both variants use the same singular-value descriptors, so this gap cannot be attributed to the use of spectral information *per se*; it is specifically the modeling of how filter sub-spaces in adjacent layers couple that

yields the improvement. This directly supports our central claim that the importance of a filter is better measured by its role in the inter-layer spectral interaction than by its isolated magnitude.



(a) Effect of inter-layer spectral coupling (b) Effect of the global adaptive budget

Figure 5. Ablation study on ResNet-56/CIFAR-10 at a matched FLOP reduction. (a) Effect of the inter-layer spectral coupling. (b) Effect of the global adaptive budget.

Effect of the global adaptive budget. Replacing FSE's global budget with a uniform layer-wise allocation; the same pruning ratio in every eligible layer, with filters still removed cheapest-first within each layer, lowers Top-1 accuracy from 93.19% to 92.96% (-0.23 pp, Fig. 5b). Since the per-filter cost is identical in both runs, this gap measures the value of allocating the pruning budget globally: redistributing capacity across layers-pruning more aggressively where filters are spectrally redundant and sparing layers where they are not-outperforms a fixed uniform schedule at matched FLOPs.

Summary. The two components are complementary and both contribute positively, with the inter-layer coupling being the larger single contributor (-0.36 pp when removed) and the global budget providing a further, independent gain (-0.23 pp when removed). Together, they account for the accuracy advantage of FSE, confirming that neither the spectral descriptor alone nor a naive global ranking is sufficient; the gain stems from their combination.

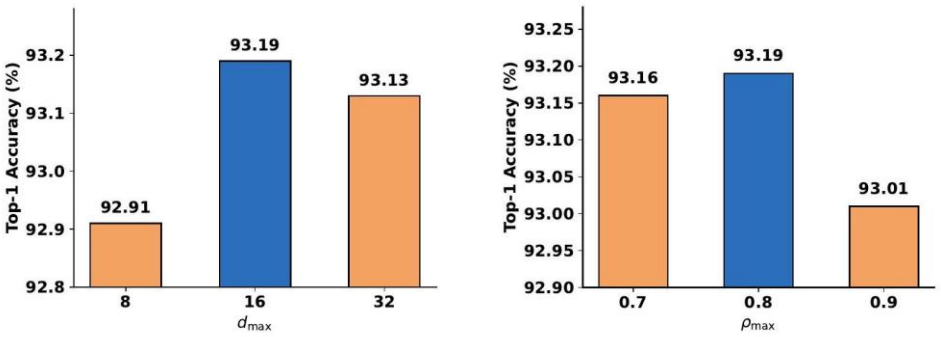
4.4 Pruning-decision Computational Overhead

The measured pruning overhead reported in Table 6, obtained on ResNet-56/CIFAR-10 at a matched FLOPs reduction of 59.84%, shows that the complete FSE pruning decision is obtained in only 0.940 s, excluding fine-tuning. The dominant cost is the construction of the SVD-based filter descriptors, which accounts for approximately 73.9% of the total time (694.73ms). This cost arises from computing the compact spectral representation of the learned filters and is incurred only once during the pruning decision. The subsequent construction of the coupling matrices requires 132.42ms (14.1%), while the SVD of the coupling matrices takes only 50.15ms (5.3%). The spectral-cost computation and global budget allocation are comparatively negligible, requiring 48.35ms (5.1%) and 14.83ms (1.6%), respectively. The peak CUDA memory consumption during importance scoring is only 16.0 MB, indicating that the coupling construction and its SVD do not introduce a substantial memory burden.

Table 6. FSE pruning overhead during importance scoring only. Fine-tuning time is excluded.

Pruning-decision component	Time (ms)
SVD filter descriptors (Eqs. 3-7)	694.73
Build coupling matrices (Eq. 10)	132.42
SVD of coupling matrices (Eq. 11)	50.15
Spectral costs (Eq. 12)	48.35
Global budget + ranking (Eqs. 13-15)	14.83
Total pruning-decision time	940.48 (0.940 s)
Peak CUDA memory during scoring	16.0 MB

Overall, these measurements show that although FSE involves additional spectral computations compared with simple criteria, such as weight norms, the complete pruning-decision stage remains lightweight, requiring less than one second and without any data-dependent forward passes. Importantly, this overhead is incurred only once before the subsequent fine-tuning stage and therefore does not constitute a recurring training cost.



(a) Sensitivity to the descriptor dimension $d_{\max}$ (b) Sensitivity to the per-layer pruning cap $\rho_{\max}$

Figure 6. Sensitivity analysis of FSE on ResNet-56/CIFAR-10 at a matched FLOP reduction. (a) Effect of the descriptor projection dimension $d_{\max}$. (b) Effect of the per-layer pruning cap $\rho_{\max}$.

4.5 Sensitivity to the Descriptor Dimension and Per-layer Cap

FSE exposes two design hyper-parameters: the descriptor projection dimension $d_{\max}$ in Eq. (8) and the per-layer pruning cap $\rho_{\max}$ in Eq. (14). We assess their influence by varying one hyper-parameter at a time at a matched FLOP reduction of 59.84%. The results are summarized in Figure 6.

Descriptor dimension $d_{\max}$. Top-1 accuracy varies by 0.28 pp across the tested range and reaches its highest value at $d_{\max} = 16$. Reducing the dimension to $d_{\max} = 8$ lowers accuracy to 92.91% (-0.28 pp), indicating that a more aggressive projection can discard information relevant to the spectral descriptors and their coupling. Increasing the dimension to $d_{\max} = 32$ does not improve the result, yielding 93.13% (-0.06 pp) relative to $d_{\max} = 16$. This suggests that increasing the descriptor dimension beyond 16 provides limited practical benefit under the tested pruning setting. Moreover, because $d_{\max}$ only affects the projected descriptor representation, its influence is limited to layers for which the original descriptor dimension exceeds the selected projection dimension. Overall, $d_{\max} = 16$ provides the best accuracy among the tested values without introducing an unnecessarily large descriptor representation.

Per-layer cap $\rho_{\max}$. The tested values produce a narrow accuracy range of 0.18 pp, with the highest accuracy obtained at $\rho_{\max} = 0.8$. Tightening the cap to $\rho_{\max} = 0.7$ results in only a 0.03 pp decrease (93.16%), indicating that a moderately stricter constraint has little effect on the final accuracy at the considered FLOP target. In contrast, increasing the cap to $\rho_{\max} = 0.9$ reduces accuracy to 93.01% (-0.18 pp). This result suggests that allowing substantially more aggressive pruning within individual layers can be less favorable, despite the same overall FLOP reduction, as it permits greater concentration of pruning in selected layers. The results therefore support $\rho_{\max} = 0.8$ as a balanced operating point between global allocation flexibility and protection against excessive layer-wise pruning.

Summary. Across the tested values, the maximum variation in Top-1 accuracy is 0.28 pp for $d_{\max}$ and 0.18 pp for $\rho_{\max}$. The relatively small variations indicate that FSE is reasonably robust to both hyper-parameters within the evaluated ranges. The selected values, $d_{\max} = 16$ and $\rho_{\max} = 0.8$, provide the best accuracy among the tested configurations while avoiding unnecessarily large descriptor dimensions or overly permissive layer-wise pruning.

5. CONCLUSION

This paper presented Filter Spectral Efficiency (FSE), a structured filter-pruning framework that estimates filter importance through the spectral coupling between consecutive convolutional layers. Unlike conventional criteria that evaluate filters independently or rely on activation statistics, FSE derives pruning decisions directly from the learned weights by modeling inter-layer information flow. The proposed framework further combines this relational spectral criterion with a global adaptive pruning budget that allocates sparsity automatically across the network while preserving structural consistency. Extensive experiments on five CNN architectures for CIFAR-10 and ResNet-50 on ImageNet demonstrated that FSE consistently achieves a favorable accuracy-compression trade-off, matching or outperforming representative structured pruning methods while requiring no calibration data or feature-map collection during the pruning stage. Ablation studies further confirmed that both the

inter-layer spectral criterion and the global adaptive budget contribute independently to the overall performance, validating the design choices of the proposed framework. These observations also suggest promising future directions for combining FSE with complementary compression techniques, such as tensor decomposition and knowledge distillation, as discussed in the positioning of this work. Future work will investigate extending the proposed spectral coupling formulation to transformer architectures and other modern vision models, incorporating hardware-aware objectives, such as latency and energy consumption, into the global budget and exploring more general forms of inter-layer dependency modeling for efficient neural network compression.

A. THEORETICAL PROPERTIES

This appendix collects the formal properties of the FSE construction. Unless stated otherwise, $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ denotes the thin singular value decomposition of a coupling matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, with $\mathbf{U} \in \mathbb{R}^{m\times r}$ satisfying $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_r$ and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1 \geq \cdots \geq \sigma_r \geq 0)$, where $r = \mathrm{rank}(\mathbf{M})$.

Lemma 1 (Information-preserving alignment). Let $\mathbf{c} \in \mathbb{R}^d$ and let $\tilde{\mathbf{c}} = [\mathbf{c}^\top, \mathbf{0}^\top]^\top \in \mathbb{R}^{d'}$ with $d' \geq d$ be its zero-padding. Then $\|\tilde{\mathbf{c}}\|_2 = \|\mathbf{c}\|_2$ and for any two descriptors padded to a common dimension the cosine similarity in Eq. (10) is independent of the padding target d'. Consequently the entries of $\mathbf{M}^{(l)}$ do not depend on the alignment dimension $d_\star$.

Proof. Appending zero coordinates leaves the sum of squares unchanged, so $\|\tilde{\mathbf{c}}\|_2^2 = \sum_j c_j^2 + 0 = \|\mathbf{c}\|_2^2$. For descriptors $\mathbf{a} \in \mathbb{R}^{d_a}$ and $\mathbf{b} \in \mathbb{R}^{d_b}$ padded to any $d' \geq \max(d_a, d_b)$, the inner product is $\langle \tilde{\mathbf{a}}, \tilde{\mathbf{b}} \rangle = \sum_{j=1}^{\min(d_a,d_b)} a_j b_j$, since in every coordinate beyond $\min(d_a, d_b)$ at least one factor is zero; this value does not depend on d'. As both norms are likewise independent of d', so is the cosine $\langle \tilde{\mathbf{a}}, \tilde{\mathbf{b}} \rangle / (\|\tilde{\mathbf{a}}\|\|\tilde{\mathbf{b}}\|)$. Truncation, by contrast, would drop nonzero terms from the inner product and is therefore not informationpreserving.

Lemma 2 (Boundedness of the coupling spectrum).

$$\sigma_1 \leq \|\mathbf{M}\|_F \leq \sqrt{mn}, \sum_{k=1}^{r} \sigma_k = \|\mathbf{M}\|_* \leq \sqrt{r}\|\mathbf{M}\|_F \leq \sqrt{rmn}.$$

Proof. By definition $M_{ij} = \cos(\mathbf{d}_i, \mathbf{d}_j)$, hence $|M_{ij}| \leq 1$ by the Cauchy-Schwarz inequality and $\|\mathbf{M}\|_F^2 = \sum_{i,j} M_{ij}^2 \leq mn$. The spectral norm satisfies $\sigma_1 = \|\mathbf{M}\|_2 \leq \|\mathbf{M}\|_F$ and applying Cauchy-Schwarz to the vector of singular values gives $\sum_k \sigma_k \leq \sqrt{r}(\sum_k \sigma_k^2)^{1/2} = \sqrt{r}\|\mathbf{M}\|_F$. Combining these yields the stated bounds.

Proposition 1 (Spectral interpretation of the pruning cost). Let $\kappa_i = \sum_k U_{ik}^2 \sigma_k$ be the pruning cost defined in Eq. (12). Then

(i) $\boldsymbol{\kappa} = \mathrm{diag}((\mathbf{M}\mathbf{M}^\top)^{1/2})$, where $(\mathbf{M}\mathbf{M}^\top)^{1/2}$ is the unique positive semidefinite square root of $\mathbf{M}\mathbf{M}^\top$.

(ii) $\sum_i \kappa_i = \sum_k \sigma_k = \|\mathbf{M}\|_*$, that is, the pruning costs exactly decompose the nuclear norm of the coupling matrix.

(iii) For every filter, $0 \leq \kappa_i \leq \sigma_1$.

Proof. (i) Since $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$,

$$\mathbf{M}\mathbf{M}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top.$$

Define $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$. The matrix $\mathbf{A}$ is symmetric and positive semidefinite and

$$\mathbf{A}^2 = \mathbf{U}\mathbf{\Sigma}(\mathbf{U}^\top\mathbf{U})\mathbf{\Sigma}\mathbf{U}^\top = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top = \mathbf{M}\mathbf{M}^\top.$$

By uniqueness of the positive semidefinite square root, $\mathbf{A} = (\mathbf{M}\mathbf{M}^\top)^{1/2}$. Its diagonal entries satisfy $A_{ii} = \sum_k U_{ik}\sigma_k U_{ik} = \sum_k U_{ik}^2\sigma_k = \kappa_i$, which proves (i).

(ii) Taking the trace, $\sum_i \kappa_i = \mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top) = \mathrm{tr}(\mathbf{\Sigma}\mathbf{U}^\top\mathbf{U}) = \mathrm{tr}(\mathbf{\Sigma}) = \sum_k \sigma_k = \|\mathbf{M}\|_*$.

(iii) Let $w_k = U_{ik}^2 \geq 0$. Since $\mathbf{U}\mathbf{U}^\top$ is an orthogonal projector, $\sum_k w_k = (\mathbf{U}\mathbf{U}^\top)_{ii} \leq 1$. Hence $0 \leq \kappa_i = \sum_k w_k\sigma_k \leq \sigma_1 \sum_k w_k \leq \sigma_1$, which completes the proof.

Remark 1 (Sign invariance). The descriptor of Eq. (5) is defined only up to sign and Eq. (6) selects one representative. This choice does not affect the criterion. Let $\boldsymbol{S}_l = \mathrm{diag}\left(\varepsilon_0^{(l)}, \ldots, \varepsilon_{C_{\mathrm{out}}^{(l)}-1}^{(l)}\right)$ with $\varepsilon_i^{(l)} \in \{-1, +1\}$ be an arbitrary sign choice for layer l. Because the projection in $Eq.$ (8) is an uncentred SVD projection, replacing $\boldsymbol{D}^{(l)}$ by $\boldsymbol{S}_l\boldsymbol{D}^{(l)}$ leaves the retained principal subspace unchanged-indeed $\boldsymbol{D}^{(l)\top}\boldsymbol{S}_l$ has the same left singular vectors $\boldsymbol{U}_D$ and the same singular values as $\boldsymbol{D}^{(l)\top}$-and yields $\boldsymbol{S}_l\widetilde{\boldsymbol{D}}^{(l)}$. Zeropadding and row-wise ℓ_2 normalisation commute with $\boldsymbol{S}_l$ as well, so $\boldsymbol{M}^{(l)} \mapsto \boldsymbol{S}_l\boldsymbol{M}^{(l)}\boldsymbol{S}_{l+1}$.

Hence $\boldsymbol{M}\boldsymbol{M}^\top \mapsto \boldsymbol{S}_l\boldsymbol{M}\boldsymbol{M}^\top\boldsymbol{S}_l$ and, by uniqueness of the positive semidefinite square root, $(\boldsymbol{M}\boldsymbol{M}^\top)^{1/2} \mapsto \boldsymbol{S}_l(\boldsymbol{M}\boldsymbol{M}^\top)^{1/2}\boldsymbol{S}_l$. Conjugation by a diagonal sign matrix preserves diagonal entries, so Proposition 1(i) gives $\kappa_i^{(l)}$ unchanged for every filter i and every layer l. The pruning costs, the induced global ranking and the resulting pruned set are therefore independent of the sign convention.

Remark 2. Proposition 1 provides a direct spectral interpretation of the proposed pruning criterion. The quantity $(\mathbf{U}\mathbf{U}^\top)_{ii}$ represents the spectral leverage of filter i, while κ_i measures the contribution of that filter to the nuclear norm of the inter-layer coupling matrix. Consequently, filters with small values of κ_i contribute only weakly to the dominant inter-layer coupling modes and are therefore natural candidates for pruning.

Proposition 2 (Optimality of the global budget). Let $\mathcal{E} = \bigcup_l \mathcal{E}_l$ be the eligible filters partitioned by layer, with per-layer caps cap_l and let the costs $\{\kappa_i^{(l)}\}$ be arbitrary reals. The greedy rule that scans the filters in nondecreasing cost and admits a filter if and only if its layer has not yet reached cap_l, stopping after T admissions, returns a set $\mathcal{P}$ that (a) is feasible, i.e. $|\mathcal{P} \cap \mathcal{E}_l| \leq \mathrm{cap}_l$ for all l and (b) has minimum total cost among all feasible sets of cardinality $|\mathcal{P}|$.

Proof. The feasible pruning sets form a partition matroid whose independent sets are $\mathcal{I} = \{S \subseteq \mathcal{E} : |S \cap \mathcal{E}_l| \leq \mathrm{cap}_l, \forall l\}$.

The proposed global allocation is exactly the classical greedy algorithm on this partition matroid: filters are processed in non-decreasing pruning cost and selected whenever the corresponding layer capacity is not exceeded. By the fundamental optimality theorem for matroid greedy algorithms [41], the resulting set is a minimum-cost feasible subset for every attainable cardinality. With the stopping rule, the attained cardinality equals $\min(T, \sum_l \mathrm{cap}_l)$, so the selected pruning set $\mathcal{P}$ is both feasible and globally optimal under the prescribed layer-wise constraints.

Proposition 3 (Dimensional consistency of the structural surgery). Consider the supported architecture class in which all producer-consumer

dependencies are correctly identified, including sequential, residual and concatenation-based topologies. If, for every pruned layer l, the output channels indexed by $\mathcal{P}_l$ are removed from $\mathbf{W}^{(l)}$ and the same indices are removed from the input channels of every consumer of layer l (Eqs. (16)-(18)), then the pruned network admits a valid forward pass on any input of the original spatial resolution and every residual addition remains dimensionally defined.

Proof. The proof follows by induction over the computation graph in topological (data-flow) order. The first convolution is unchanged and therefore preserves the original input dimension. Assume that all predecessors of layer l satisfy the required channel consistency. Removing the output channels indexed by $\mathcal{P}_l$ from layer l and the corresponding input channels from every consumer preserves dimensional agreement along every producer-consumer edge; the accompanying batch-normalization parameters are indexed by the same retained set $\mathcal{K}_l$ and hence remain consistent. Residual additions remain valid because the protected branches in $\mathcal{S}$ preserve their original output dimensions, while concatenation operators receive tensors whose channel dimensions are consistently updated. Therefore every producer-consumer dependency remains dimensionally consistent, thereby guaranteeing that the structurally pruned network admits a valid forward pass.

REFERENCES

[1] E. Edozie et al., "Comprehensive Review of Recent Developments in Visual Object Detection Based on Deep Learning," Artificial Intelligence Review, vol. 58, no. 9, p. 277, 2025.

[2] M. H. Reza et al., "A Comprehensive Review of Convolutional Neural Networks: Foundations, Enhancements and Applications," Neural Computing and Applications, vol. 38, no. 4, p. 56, 2026.

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[4] G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," Proc. of the IEEE Conf. on Comp. Vision and Patt. Recog. (CVPR), pp. 2261-2269, 2017.

[5] C. Szegedy et al., "Going Deeper with Convolutions," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1-9, Boston, USA, 2015.

[6] H. Cheng, M. Zhang and J. Q. Shi, "A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis and Recommendations," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, pp. 10558-10578, 2024.

[7] R. Schwartz, J. Dodge, N. A. Smith and O. Etzioni, "Green ai," Commun. ACM, vol. 63, no. 12, p. 54-63, [Online], Available: https://doi.org/10.1145/3381831, Nov. 2020.

[8] J. Yang et al., "A Survey on Neural Network Quantization," Proc. of the 2025 6th Int. Conf. on Computer Infor. and Big Data Appl. (CIBDA '25), Association for Comp. Machinery, p. 384-394, NY, USA, 2025.

[9] S. Zhang and R. Saab, "Theoretical Guarantees for Low-rank Compression of Deep Neural Networks," Applied and Computational Harmonic Analysis, vol. 82, p. 101837, 2026.

[10] T. Shehzadi et al., "Knowledge Distillation in Object Detection: A Survey from CNN to Transformer," Sensors (Basel, Switzerland), vol. 26, no. 1, p. 292, 2026.

[11] K. Zhu et al., "A Comprehensive Review of Network Pruning Based on Pruning Granularity and Pruning Time Perspectives," Neurocomputing, vol. 626, p. 129382, 2025.

[12] S. Han, J. Pool, J. Tran and W. J. Dally, "Learning both weights and connections for efficient neural networks," in Advances in Neural Information Processing Systems (NeurIPS), 2015, pp. 1135-1143.

[13] H. Li, A. Kadav, I. Durdanovic, H. Samet and H. P. Graf, "Pruning Filters for Efficient Convnets," Proc. of the Int. Conf. on Learning Representations (ICLR), arXiv: 1608.08710, 2017.

[14] Y. He, P. Liu, Z. Wang, Z. Hu and Y. Yang, "Filter Pruning *via* Geometric Median for Deep Convolutional Neural Networks Acceleration," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4340-4349, 2019.

[15] M. Lin etal., "HRank: Filter Pruning Using High-rank Feature Map," Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1529-1538, 2020.

[16] S. Lin et al., "Towards Optimal Structured CNN Pruning *via* Generative Adversarial Learning," Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2785-2794, 2019.

[17] T. Yuan, S. Yang, B. Liu and Y. Tang, "Explainable CNN Filter Pruning Based on Shapley Value Approximation in View of Game Theory," Expert Systems with Appl., vol. 300, p. 130232, 2025.

[18] E. Chen, H. Wang, Z. Shi and W. Zhang, "Channel Pruning for Convolutional Neural Networks Using l0-Norm Constraints," Neurocomputing, vol. 636, p. 129925, 2025.

[19] Ambuj, "Feature Representation Fidelity Preservation During Neural Network Pruning for Enhanced Compression Efficiency," Neurocomputing, vol. 634, p. 129901, 2025.

[20] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," Uni. of Toronto, Technical Report, [Online], Available: https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf, 2009.

[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," Proc. of the Int. Conf. on Learning Representations (ICLR), 2015.

[22] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," Int. Journal of Computer Vision, vol. 115, no. 3, pp. 211-252, 2015.

[23] Y. Sui et al., "Chip: Channel Independence-based Pruning for Compact Neural Networks," Advances in Neural Information Processing Systems, vol. 34, pp. 24604-24616, 2021.

[24] M. Lin, L. Cao, S. Li, Q. Ye, Y. Tian, J. Liu, Q. Tian and R. Ji, "Filter Sketch for Network Pruning," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 7091-7100, 2021.

[25] J.-H. Luo et al., "ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression," Proc. of the IEEE Int. Conf. on Computer Vision, pp. 5058-5066, Venice, Italy, 2017.
[26] Y. He, X. Zhang and J. Sun, "Channel Pruning for Accelerating Very Deep Neural Networks," Proc. of the IEEE Int. Conf. on Computer Vision, pp. 1389-1397, Venice, Italy, 2017.
[27] M. Alwani et al., "DECORE: Deep Compression with Reinforcement Learning," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 12349-12359, New Orleans, USA, 2022.
[28] Y. Zheng et al., "A Novel and Efficient Model Pruning Method for Deep Convolutional Neural Networks by Evaluating the Direct and Indirect Effects of Filters," Neurocomputing, vol. 569, p. 127124, 2024.
[29] Y. Lian, P. Peng, K. Jiang and W. Xu, "Cross-layer Importance Evaluation for Neural Network Pruning," Neural Networks, vol. 179, p. 106496, 2024.
[30] Y. He et al., "Soft Filter Pruning for Accelerating Deep Convolutional Neural Networks," Proc. of the 27th Int. Joint Conf. on Artificial Intelligence (IJCAI), pp. 2234-2240, 2018.
[31] V. T. Pham, Y. Zniyed and T. P. Nguyen, "Efficient Tensor Decomposition-based Filter Pruning," Neural Networks, vol. 178, p. 106393, 2024.
[32] C. Zhen, W. Zhang, J. Mo, M. Ji, H. Zhou and J. Zhu, "Rasp: Regularization-based Amplitude Saliency Pruning," Neural Networks, vol. 168, pp. 1-13, 2023.
[33] Z. Liu et al., "Learning Efficient Convolutional Networks through Network Slimming," Proc. of the 2017 IEEE Int. Conf. on Computer Vision (ICCV), pp. 2755-2763, Venice, Italy, 2017.
[34] M. Lin, R. Ji, Y. Zhang, B. Zhang, Y. Wu and Y. Tian, "Channel Pruning *via* Automatic Structure Search," Proc. of the 29th Int. Joint Conf. on Artificial Intelligence (IJCAI'20), pp. 673-679, 2021.
[35] Y. Xue, W. Yao, S. Peng and S. Yao, "Automatic Filter Pruning Algorithm for Image Classification," Applied Intelligence, vol. 54, no. 1, pp. 216-230, 2024.
[36] V. T. Pham et al., "Enhanced Network Compression through Tensor Decompositions and Pruning," IEEE Trans. on Neural Networks and Learning Systems, vol. 36, no. 3, pp. 4358-4370, 2025.
[37] V. T. Pham et al., "Coupled Tensor Decomposition for Compact Network Representation," IEEE Trans. on Neural Networks and Learning Systems, vol. 37, no. 2, pp. 617-631, 2026.
[38] W. Lin et al., "STPP: Efficient and Progressive Structured Pruning *via* Enhanced Sparsification Paradigm," IEEE Trans. on Pattern Analysis and Machine Intelligence, Early Publishing, pp. 1-17, 2026.
[39] G. Fang et al., "DepGraph: Towards Any Structural Pruning," Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 16091-16 101, Vancouver, Canada, 2023.
[40] Z. Huang and N. Wang, "Data-driven Sparse Structure Selection for Deep Neural Networks," Proc. of the 15th Europ. Conf. on Comp. Vision (ECCV 2018), Part 16, pp. 317-334, Munich, Germany, 2018.
[41] A. Schrijver, Combinatorial Optimization: Polyhedra and Efficiency, E-ISSN: 2197-6783, Springer, 2003.

ملخص البحث:

يُعدّ تقليم المرشّحات المهيكل نهجاً فعّالاً لتقليل التّكلفة الحسابية للشّبكات العصبية العميقة مع الحفاظ على إمكانية النّشر الفعّال على العتاد القياسي. ومع ذلك، فإنّ معظم طرق التّقليم الحالية تقيم المرشّحات بشكلٍ مستقلّ، متجاهلةً الاقتران الطّيفي بين الطّبقات المتتالية، في حين تعتمد الأساليب عالية الأداء على إحصائيات التّنشيط أو بيانات المعايرة.

نقترح في هذه الورقة "الكفاءة الطّيفية للمرشّحات" عبر إطار عملٍ للتّقليم المهيكل لا يعتمد على البيانات، حيث يقدّر أهمّية المرشّح من خلال الاقتران الطّيفي بين الطّبقات المتجاورة. ويعمل الإطار المقترح على إنشاء واصِفاتٍ للمرشّحات تعتمد على القيم المفردة، وبناء مصفوفات الاقتران بين الطّبقات، واشتقاق معيارٍ للتّقليم الطّيفي، وتوزيع التّناثُر عبر ميزانية تكييفية شاملة.

وقد أظهرت التّجارب المكثّفة التي أجريت على خمس بنياتٍ للشّبكات العصبية التّلافيفية أنّ الإطار المقترح يحقّق باستمرارٍ توازناً مُمتازاً بين الدّقّة ومعدّل الضّغط، متوّفقاً بذلك على الطّرق النّموذجية للتّقليم المهيكل عند مستوياتٍ مماثلةٍ من تقليل المعْلمات والعمليات الحسابية. وبينت النّتائج أنّ الرّبْط الطّيفي بين الطّبقات يوفّر أساساً فعّالاً وعملياً لتّقليم الشّبكات المهيكل دون الحاجة إلى بياناتٍ.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).