Topology-Weighted Effective Rank: A Zero-Cost Proxy for Training Dynamics Stability in Neural Architecture Search

Haojie Zhang[1], Yeming Yang[1], Songbai Liu[1⋆], Lijia Ma[1], Ka-Chun Wong[2], and Qiuzhen Lin[1]

[1] Shenzhen University, Shenzhen, China
[2] City University of Hong Kong, Hong Kong, China

Abstract. Recent training-free Neural Architecture Search (NAS) methods have introduced zero-cost proxies (ZCPs) to automate architecture design without expensive training or expert intervention. However, existing ZCPs either completely ignore training dynamics or rely on static or initialization-time characterizations of optimization that fail to capture fine-grained correlations with performance. Moreover, they overlook the heterogeneous importance of different model components, which limits their ability to generalize across datasets and search spaces. To mitigate these limitations, we propose an Effective Rank Score called **ER-Score** as a novel ZCP to quantify **stability in the training dynamics of over-parameterized networks**. Experiments show that ER-Score consistently outperforms existing ZCPs across multi-scale tasks and diverse architectures, including convolutional neural networks (CNNs) and Vision Transformers (ViTs). We further introduce topology-weighted strategies for CNNs that incorporate topological information via operation-wise feature map aggregation, resulting in a topology-weighted variant termed **TER-Score**. Extensive experiments demonstrate that TER-Score achieves **state-of-the-art** ranking consistency and the best population convergence when integrated with evolutionary search on the NAS-Bench-301 benchmark. Finally, our method achieves the **lowest test errors of 2.41% and 23.55%** on the DARTS search space for the CIFAR-10 and ImageNet-1k datasets, respectively. Code is released at `https://github.com/Thiswycf/TER-Score`.

Keywords: Neural Architecture Search · Training-Free NAS · Zero-cost proxy · Training Dynamics

1 Introduction

The design of deep neural networks (DNNs) has relied on extensive expert knowledge and labor-intensive manual tuning, substantially limiting scalability and real-world deployment [6, 18, 22]. To overcome this, neural architecture search (NAS) automatically discovers high-performing architectures, greatly reducing

⋆ Corresponding author: `songbai@szu.edu.cn`

human effort and computational costs [2, 48]. One-shot NAS methods [16, 27, 43] have leveraged weight-sharing to improve efficiency, avoiding the prohibitive expense of early NAS [2]. However, despite these gains, they still incur non-negligible training overhead, making them unsuitable for large-scale or resource-constrained settings. In contrast, zero-cost proxies (ZCPs) offer a promising alternative by enabling training-free performance estimation while maintaining reasonable correlation with fully trained accuracies [17, 21, 23, 31, 39, 44].

Despite the growing popularity of ZCPs, existing methods remain limited in terms of training dynamics. Most ZCPs rely on static architectural or initialization-dependent signals, such as parameter statistics or single-pass forward or backward measurements, which ignore how optimization behavior evolves during training [1, 17, 24, 31, 39]. Other proxies attempt to incorporate optimization-related signals but typically fail to model the dynamics during training [6,21,23]. In practice, the performance of neural networks is mainly affected by training dynamics, including convergence behavior, optimization stability, and sensitivity to perturbations [13]. As a result, existing ZCPs often exhibit systematic inconsistencies in architecture ranking across different tasks or search spaces.

Furthermore, many existing ZCPs [1, 17, 21] adopt a node-homogeneous perspective of neural architectures, where global scores are obtained by naively aggregating local statistics across layers or nodes (e.g., summation [17, 19, 21, 23] or averaging [21, 24]). This simplification overlooks the fact that architectural components contribute unequally to optimization behavior, as their roles vary with depth, connectivity patterns, and functional semantics [4, 10, 35, 40].

To address the above limitations, we introduce an Effective Rank Score (ER-Score) as a novel ZCP that evaluates neural architectures from the perspective of stability in training dynamics. ER-Score provides a principled criterion for training-free architecture comparison by characterizing optimization behavior rather than static architectural signals. Moreover, to account for the structural heterogeneity of convolutional neural networks (CNNs), we incorporate topology-aware weighting mechanisms into ER-Score. This results in a topology-weighted variant, termed TER-Score, which explicitly models the relative importance of architectural components induced by network topology.

As an overview, Figure 1 summarizes the average ranking performance of ER-Score and TER-Score across multiple NAS benchmarks and computer vision tasks, providing preliminary empirical evidence for our approach's effectiveness.

Our main contributions are summarized as follows:

- We identify **stability in training dynamics** as a key factor for reliable zero-cost evaluation of neural architectures and propose a principled ZCP, termed **ER-Score**, to characterize this property in a training-free manner.
- We introduce a **topology-weighted** extension, called **TER-Score**, which incorporates architectural topology into the evaluation process and models the heterogeneous roles of components in CNNs.
- We demonstrate through extensive experiments that modeling stability in training dynamics and architectural topology leads to more consistent architecture ranking and more reliable guidance for downstream NAS algorithms.

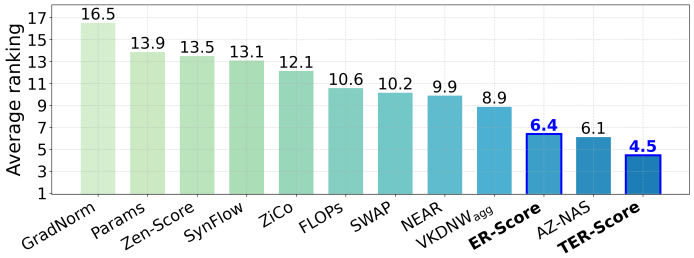


Fig. 1: Average performance ranking across ten computer vision tasks in three NAS benchmarks [12, 14, 45], where lower ranks indicate better overall performance.

2 Related Work

2.1 Training Dynamics and Trainability-Based Proxies

Several studies have investigated architecture evaluation from the perspective of training dynamics and trainability. TE-NAS [6] analyzes neural architectures in the infinite-width regime and studies optimization behavior through neural tangent kernel (NTK) theory. However, in practice, its trainability evaluation is instantiated by evaluating the condition number of the NTK at initialization, which fails to capture the evolution of training dynamics over time. ZiCo [23] establishes a theoretical connection between gradient compactness at initialization and network performance. While effective in revealing a proportional relationship between early optimization behavior and performance, ZiCo only captures coarse-grained characteristics of the optimization landscape and cannot further refine the features in the training dynamics. AZ-NAS [21] embeds trainability-related signals into a nonlinear ranking framework together with a set of architectural proxies. Although this combination yields strong empirical performance, the role of trainability is mainly supported by practical observations rather than theoretical analysis, and the trainability signals exhibit limited consistency when considered individually. To alleviate the limitations, our work adopts stability in training dynamics as a principled criterion for zero-cost architecture evaluation.

2.2 Structural Awareness in Architecture Evaluation

Another line of research has recognized that neural architectures are structurally heterogeneous, and different components may contribute unequally to performance depending on depth, connectivity, and functional semantics [4, 35, 36, 40]. However, most existing ZCPs adopt a node-homogeneous evaluation paradigm, where node-wise statistics are uniformly aggregated to produce a global architecture score [1, 17, 21, 23]. Recent work has explicitly questioned the node-homogeneous assumption in zero-cost evaluation. Dong et al. [10] show that node-wise zero-cost statistics contribute unequally to performance estimation and propose a simple sinusoidal weighting strategy inspired by the Mixer Architecture with Bayesian Network. In contrast, our approach assigns topology-aware

weights, providing an explicit and interpretable representation of component importance and significantly improving performance.

3 Methodology

In this section, we present our zero-cost architecture evaluation framework, which is grounded in the theory of training dynamics. We introduce the Effective Rank Score (ER-Score), a training-free metric designed to characterize optimization behavior rather than static architectural properties.

To account for the structural heterogeneity of CNNs, we model a CNN as a directed acyclic graph (DAG) and propose four topology-aware weighting strategies to differentiate the relative importance of individual modules. By integrating ER-Score with these topology-aware weights, we derive the topology-weighted Effective Rank Score, termed TER-Score, which incorporates structural information into training-free architecture evaluation and provides a principled criterion.

3.1 Preliminaries

Notations We define $[n] = \{1, 2, \ldots, n\}$. We denote scalars by lowercase letters (e.g., x), vectors by bold lowercase letters (e.g., $\mathbf{x}$), and matrices by bold uppercase letters (e.g., $\mathbf{X}$). Given an event $\mathcal{A}$, let $\mathbb{I}\{\mathcal{A}\}$ denote the indicator function of whether $\mathcal{A}$ occurs. For a matrix $\mathbf{A}$, $\mathbf{A}_{ij}$ denotes its (i, j)-th entry. We define the training set as $\mathbf{X} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^{d\times 1}, y_i \in \mathbb{R}, i \in [n]\}$, where $\mathbf{x}_i$ and y_i represent the i-th input and label, respectively. $\mathcal{N}(0, \mathbf{I})$ and $\mathcal{U}\{S\}$ represent the standard Gaussian distribution and uniform distribution over a set S, respectively. The transpose of a matrix $\mathbf{X}$ is denoted by $\mathbf{X}^\top$. $\|\mathbf{X}\|_2$ denotes the matrix ℓ_2-norm, i.e., the largest singular value of $\mathbf{X}$. For a symmetric positive semi-definite matrix $\mathbf{A}$, $\mathrm{Tr}(\mathbf{A})$ and $\mathrm{rank}(\mathbf{A})$ denote its trace and rank, respectively. We denote b, c, h, and w as the batch size, channel number, height, and width of a feature map, respectively. All logarithms mentioned below are natural.

For an input $\mathbf{x} \in \mathbb{R}^{d\times 1}$, a weight vector $\mathbf{w} \in \mathbb{R}^{d\times 1}$ in the weight matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$, and output weights $\mathbf{a} \in \mathbb{R}^{m\times 1}$, let $\mathbf{f}(\mathbf{W}, \mathbf{a}, \mathbf{x})$ denote a single-hidden-layer neural architecture (NN) defined as:

$$\mathbf{f}(\mathbf{W}, \mathbf{a}, \mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \mathbf{a}_r \sigma(\mathbf{w}_r^\top \mathbf{x}), \tag{1}$$

where $\sigma(z) = z\mathbb{I}\{z > 0\}$ is the ReLU activation function. Given the training set $\mathbf{X}$, the empirical loss is formulated as:

$$l(\mathbf{W}, \mathbf{a}) = \sum_{i=1}^{n} \frac{1}{2} \big(\mathbf{f}(\mathbf{W}, \mathbf{a}, \mathbf{x})_i - y_i\big)^2. \tag{2}$$

Following the definitions in [13], we introduce the Gram matrices $\mathbf{H}(t)$ and $\mathbf{H}^\infty$ as follows.

Definition 1 (Gram Matrix). *For a single-hidden-layer NN, the Gram matrix* $\mathbf{H}(t) \in \mathbb{R}^{n\times n}$ *induced by the ReLU activation function over the training set* $\mathbf{X}$ *is defined element-wise as:*

$$[\mathbf{H}(t)]_{ij} = \frac{1}{m}\sum_{r=1}^{m} \mathbf{x}_i^\top \mathbf{x}_j \mathbb{I}\{\mathbf{w}_r^\top(t)\mathbf{x}_i \geq 0,\, \mathbf{w}_r^\top(t)\mathbf{x}_j \geq 0\}, \tag{3}$$

where $\mathbf{w}_r(t)$ *denotes the* r*-th hidden weight vector at iteration* t*. The limiting matrix* $\mathbf{H}^\infty$ *is defined as:*

$$[\mathbf{H}^\infty]_{ij} = \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{I})}\left[\mathbf{x}_i^\top \mathbf{x}_j \mathbb{I}\{\mathbf{w}^\top \mathbf{x}_i \geq 0,\, \mathbf{w}^\top \mathbf{x}_j \geq 0\}\right]. \tag{4}$$

Training Dynamics We begin by recalling the training dynamics of NNs, following the formula in the proof of Theorem 3.2 in [13]. Let $\mathbf{u}(t) \in \mathbb{R}^n$ denote the network output vector in training iteration t, and let $\mathbf{y} \in \mathbb{R}^n$ denote the corresponding ground-truth label vector. The evolution of $\mathbf{u}(t)$ under gradient descent can be expressed as

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{H}(t)\left(\mathbf{y} - \mathbf{u}(t)\right), \tag{5}$$

Remark 1. The Gram matrix $\mathbf{H}(t)$ can capture the prediction dynamics at iteration t, providing a theoretical foundation to quantify the stability and expressivity of the network in subsequent analysis.

According to previous work [13], for sufficiently over-parameterized NNs, the Gram matrix $\mathbf{H}(t)$ remains nearly constant during training. Formally, for all $t \geq 0$, we have:

$$\|\mathbf{H}(t) - \mathbf{H}^\infty\|_2 \to 0, \quad \text{as } m \to \infty, \tag{6}$$

where m denotes the network width and $\mathbf{H}^\infty$ represents the limiting Gram matrix.

Remark 2. In the infinite-width regime, the training dynamics governed by (5) can be approximated by a linearized system with a fixed kernel $\mathbf{H}^\infty$. Thus, the evolution of the network outputs $\mathbf{u}(t)$ can be effectively characterized by the stationary Gram matrix $\mathbf{H}^\infty$, which captures the intrinsic stability of the network throughout training.

Theorem 1. *Let* $\{\mathbf{x}_i\}_{i=1}^n$ *be the training samples such that* $\|\mathbf{x}_i\|_2 = 1$ *for all* $i \in [n]$*. Consider a one-hidden-layer NN defined as in* (1)*, where the weights are initialized independently according to* $\mathbf{w}_r \sim \mathcal{N}(0, \mathbf{I})$ *and* $a_r \sim \mathcal{U}[-1, 1]$ *for* $r \in [m]$*. Assume that the network is sufficiently overparameterized* ($m \to \infty$) *and the limiting Gram matrix* $\mathbf{H}^\infty \succ 0$*. Then, the training loss dynamics of the NN follow*

$$l(t) = \frac{1}{2}\varepsilon_0^\top V \exp(-2\Lambda t) V^\top \varepsilon_0, \tag{7}$$

where $l(t) \coloneqq l(\mathbf{W}(t), \mathbf{a})$*,* $\mathbf{H}^\infty = V\Lambda V^\top$ *with* $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$*, and* $\varepsilon_0 \coloneqq \mathbf{u}(0) - \mathbf{y}$ *denotes the prediction residual at initialization.*

Remark 3. Theorem 1 indicates that the training loss decay is governed by the eigen-spectrum of the Gram matrix $\mathbf{H}^{\infty}$. Each eigenvalue λ_i corresponds to a descent direction in the loss landscape, determining the convergence rate along that direction. **A broader and more balanced spectrum indicates more stable training dynamics.**

3.2 ER-Score

However, computing $\mathbf{H}^{\infty}$ is intractable for DNNs, as it requires evaluating the infinite-width Gram matrix defined on the entire training set. To approximate this behavior within a tractable form, we follow the theoretical insight of [19], which shows that the training convergence rate and generalization capacity of an over-parameterized network are jointly governed by the minimum eigenvalue of its Gram matrix $\mathbf{H}^{\infty}$. Moreover, it is proved in [19] that the convolutional layer in a CNN can be regarded as a multi-sample fully-connected layer with constraints, where each channel of the feature map acts as an independent sample. This equivalence enables us to transfer the spectral analysis of $\mathbf{H}^{\infty}$ to the correlation structure within feature maps.

Specifically, for a feature map $\mathbf{A} \in \mathbb{R}^{b\times h\times w\times c}$ extracted from an intermediate module, we flatten it into $\mathbf{A}'$ and compute its per-sample correlation matrix:

$$\mathbf{C}_i = \frac{1}{hw}(\mathbf{A}'_i)^{\top}\mathbf{A}'_i, \quad \mathbf{A}'_i \in \mathbb{R}^{(hw)\times c}, \tag{8}$$

where $i \in [b]$, $\mathbf{C}_i$ is a symmetric positive semi-definite matrix. According to the theoretical results in [19], the eigenvalues of $\mathbf{C}_i$ reflect the local training dynamics, analogous to those characterized by the eigen-spectrum of $\mathbf{H}^{\infty}$. Therefore, the Effective Rank of $\mathbf{C}_i$ can serve as a feasible proxy for evaluating the stability and capacity of DNNs, while preserving the correspondence to the theoretically grounded analysis of $\mathbf{H}^{\infty}$.

Building upon the above motivation, we formally introduce the ER-Score to quantify stability in the training dynamics of feature representations in a neural architecture. Let $\mathbf{A}_l \in \mathbb{R}^{h_l\times w_l\times c_l}$ denote the feature map of the l-th component in the network, where h_l, w_l, and c_l are its height, width, and number of channels, respectively. We first flatten $\mathbf{A}_l$ into a two-dimensional matrix $\mathbf{A}'_l$, and compute its feature correlation matrix:

$$\mathbf{C}_l = \frac{1}{h_l w_l}(\mathbf{A}'_l)^{\top}\mathbf{A}'_l, \quad \mathbf{A}'_l \in \mathbb{R}^{(h_l w_l)\times c_l}. \tag{9}$$

Since $\mathbf{C}_l$ is symmetric positive semi-definite, it admits an eigen-decomposition $\mathbf{C}_l = \mathbf{U}_l \Lambda_l \mathbf{U}_l^{\top}$, where $\Lambda_l = \mathrm{diag}(\lambda_{l,1}, \ldots, \lambda_{l,c_l})$ contains its eigenvalues.

Definition 2. *To measure the dispersion of the eigen-spectrum, we normalize the eigenvalues as* $\tilde{\lambda}_{l,i} = \frac{\lambda_{l,i}}{\sum_{j=1}^{c_l}\lambda_{l,j}}$ *and define the* ER-Score *of* $\mathbf{A}_l$ *by the Effective Rank [33] of* $\mathbf{C}_l$ *as:*

$$\mathrm{ER}(\mathbf{A}_l) = \exp\left(-\sum_{i=1}^{c_l}\tilde{\lambda}_{l,i}\log\tilde{\lambda}_{l,i}\right). \tag{10}$$

The exponential term converts the spectral entropy into a continuous analog of matrix rank, preserving its linear composability across different components.

In Equation 10, the effective rank [33] captures how evenly information is distributed across principal directions of the feature correlation matrix, providing a scale-invariant measure of structural balance that is closely related to optimization robustness, rather than merely the magnitude of activations or gradients.

For a network composed of L modules, we aggregate the effective ranks of all component correlation matrices as follows:

$$\text{ER-Score} = \sum_{l=1}^{L} \text{ER}(\mathbf{A}_l), \tag{11}$$

Although ER-Score is computed at initialization, it is designed to serve as an observation of how architectural structure conditions subsequent optimization behavior, rather than as a static descriptor of expressivity.

To investigate the correlation between the proposed ER-Score and stability in training dynamics, we conducted an analysis on NAS-Bench-201 [12], as shown in Figure 2. We selected the top 60% of architectures ranked by ER-Score and divided them into six groups, each containing 1562 architectures, and trained for 100 steps. We report the cosine similarity standard deviation between the gradients of consecutive training steps to quantify stability in training dynamics. For each training step, we averaged the cosine similarities across all architectures within the same group, and the resulting step-wise distributions were visualized using boxplots. This analysis demonstrates that architectures with higher ER-Score tend to exhibit more consistent gradient evolution during early training.

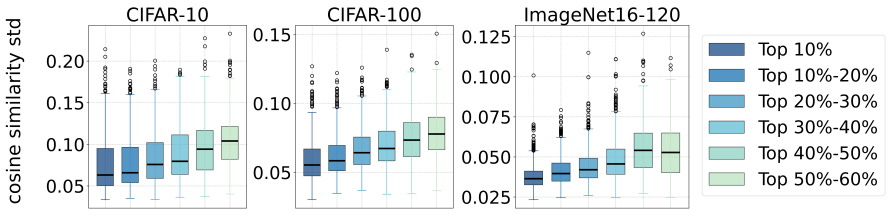


Fig. 2: Boxplots of the cosine similarity standard deviation between consecutive training steps on NAS-Bench-201 [12].

3.3 Topology-Weighting Strategy

Although the ER-Score provides a compact measure of stability in training dynamics, it inherently treats all components as equally important and ignores

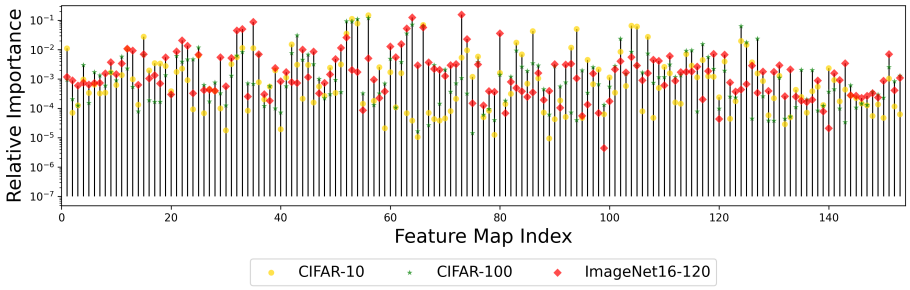


Fig. 3: Relative importance of ER-Score in feature maps based on GBDT impurity on NAS-Bench-201.

Table 1: Topological weighting formulas.

Method	**Node-level (w_v)**	**Edge-level (w_e)**
Degree	$k_v^{in} + k_v^{out}$	$k_v^{in} + k_u^{out}$
Connectivity	$c_v^s \times c_t^v$	$c_u^s \times c_t^v$
Shortest-path [9]	$((1 + d_v^s) \times (1 + d_t^v))^{-1}$	$((1 + d_u^s) \times (1 + d_t^v))^{-1}$
PageRank [29]	$\pi^{i+1} = \frac{1}{n}(1 - D) + DP^{\mathsf{T}}\pi^i$	$\pi_e^* \times \pi_e'^*$

their relative importance. Since CNN search spaces have been extensively studied [12, 14, 25, 32, 34, 45], we focus on analyzing the relative importance of CNNs.

Dong et al. [10], using gradient boosting decision trees (GBDT) [15] based on impurity regression, pointed out that node-wise ZC statistics significantly vary in their contributions to performance. We used the same regression method to perform a relative importance analysis of the operator-level feature map metric ER-Score on different tasks of NAS-Bench-201 [12], as shown in Figure 3. The results show that the contribution of operator-level feature map metric statistics to performance varies and lacks a clear distribution pattern.

This limitation motivates us to consider topological differences among components in a CNN, as the position of each node plays a distinct role in the information flow within the DAG representation of neural architectures. Accordingly, topology-weighting is introduced as an orthogonal enhancement to ER-Score to incorporate structural heterogeneity into architecture evaluation, instead of altering the underlying stability criterion.

To incorporate topological information, we propose four topology-weighting strategies, namely degree, connectivity, shortest-path, and PageRank weighting. Each strategy assigns a structural importance weight to every node, which is then propagated to edges for computing the topology-weighted ER-Score, ensuring compatibility with feature correlation computation that is naturally defined over connections between modules. Table 1 shows the main calculation formulas.

We detail four concrete instantiations of topology-weighting that operationalize the above principles as follows.

Degree-based weighting. For a DAG $G = (V, E)$, let k_v^{in} and k_v^{out} denote the in-degree and out-degree of node $v \in V$, respectively. To align with edge-level feature propagation, this weight is transferred to each edge $e = (u, v) \in E$, so that edges connecting highly interactive nodes receive greater importance.

Connectivity-based weighting. Assume the DAG G has a unique source node s and a unique sink node t. Let c_v^s denote the number of distinct paths from s to v, and c_t^v denote the number of paths from v to t. For edge-level transformation, we design the corresponding formula in Table 1 to capture the edge's contribution to the global flow of information from source to sink.

Shortest-path-based weighting. Under the same assumption of a unique source s and sink t, let d_v^s and d_t^v denote the shortest path lengths [9] from s to v and from v to t, respectively. At the edge-level, the weight emphasizes edges closer to the source or sink and more influential in information propagation.

PageRank-based weighting. To capture both aggregation and diffusion importance globally, we employ an iterative rule from PageRank [29]. Let π^i denote the node importance vector at iteration i, $n = |V|$ be the number of nodes, $D \in (0, 1)$ denote the damping factor, and P denote the column-stochastic adjacency matrix, as shown in the corresponding formula in Table 1. The stationary solution π^* gives node-level importance. For edge-level transformation, π_e^* and $\pi_e'^*$ are the PageRank scores of edge e computed on the original and reversed DAGs, respectively. The formula ensures that each edge's weight reflects both its aggregation and diffusion roles within the architecture.

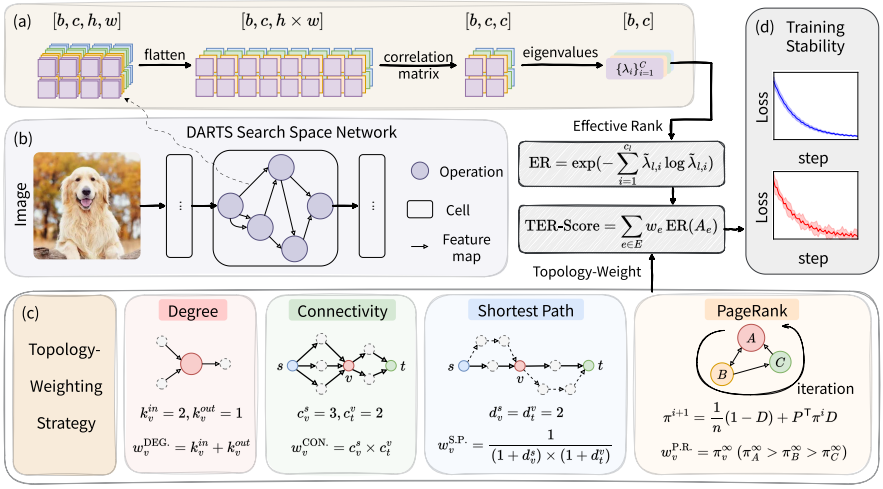


Fig. 4: Illustration of TER-Score. (a) Pipeline from a 4-dimensional feature map to build ER-Score; (b) Schematic diagram of the network architecture of DARTS [25] search space. (c) Four topology-weighting strategies producing the weight of each component to build TER-Score. (d) Schematic diagram of stability in training dynamics indicated by TER-Score.

3.4 TER-Score

Formally, the TER-Score (Topology-weighted ER-Score) is defined as:

$$\text{TER-Score} = \sum_{e \in E} w_e \, \text{ER}(A_e), \tag{12}$$

where A_e denotes the feature map corresponding to edge e and w_e is the normalized topological weight satisfying $\sum_{e \in E} w_e = 1$.

A detailed illustration of TER-Score is provided in Figure 4.

4 Experiments

We evaluate our methods on widely used NAS benchmarks with diverse search spaces, including NAS-Bench-201 (NB201) [12], NAS-Bench-301 (NB301) [45], TransNAS-Bench-101-MACRO/MICRO (TNB-A/I) [14], and ViT-Bench-101 [41]. These benchmarks cover multiple vision datasets and tasks, such as three datasets from NB201 [12]: CIFAR-10/CIFAR-100 (C10/C100) [20], ImageNet16-120 (IN) [7], seven datasets from TNB [14]: scene classification (SCE.), object classification (OBJ.), surface normal estimation (NOR.), room layout prediction (ROOM), semantic segmentation (SEG.), jigsaw puzzle recognition (JIG.), and autoencoding (A.E.), and two datasets from ViT-Bench-101 [41]: Flowers [28] and Chaoyang [47].

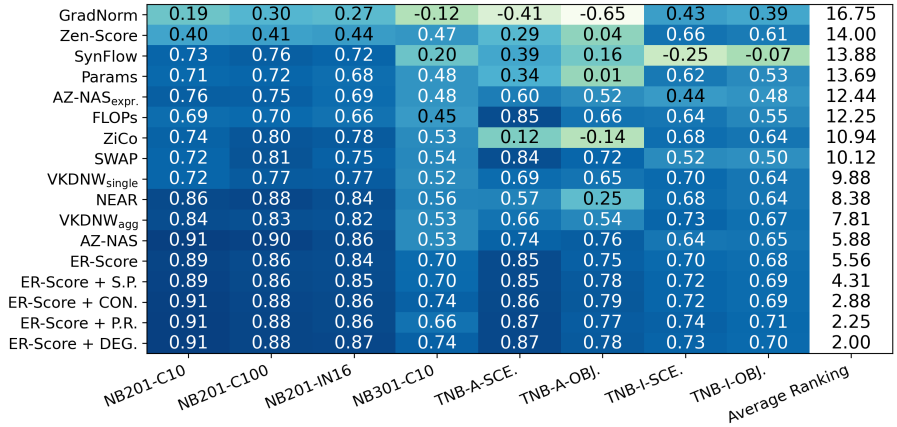


Fig. 5: Spearman's rank correlation coefficients between ZCPs and networks' ground-truth performance for 12 existing metrics and our methods on *image classification tasks*. The rows are sorted based on average ranks of independent experiments for each metric.

To quantify the consistency between proxy-based evaluations and the ground-truth performance, we adopt Spearman's rank correlation coefficient [37]. Furthermore, to evaluate the practical effectiveness of the TER-Score in architecture optimization, we integrate it into an evolutionary search framework (details are provided in the supplementary material), implemented on the DARTS search space [25]. The discovered architectures are fully trained from scratch on the C10 and ImageNet-1k [8] datasets to evaluate their performance in real NAS scenarios. All experiments were conducted using the PyTorch [30] framework on an NVIDIA RTX 4090 GPU.

4.1 Results on NAS Benchmarks

Comparison with Existing ZCPs We compare our methods with 10 existing ZCPs widely adopted in NAS research, including GradNorm [1], Zen-Score [24], SynFlow [38], ZiCo [23], SWAP [31], AZ-NAS [21], NEAR [17], and VKDNW [39], following their official implementations. #Param and #Flops are naive proxies [1]. We denote *ER-Score + DEG./CON./S.P./P.R.* as the ER-Score using weighting strategies based on *degree*, *connectivity*, *shortest-path*, and *PageRank*, respectively. We report the Spearman's rank correlation coefficients on NB201, NB301, and TNB-A/I benchmarks, covering both image classification tasks and downstream transfer tasks, as shown in Figures 5 and 6, respectively. Furthermore, we also tested the general applicability of ER-Score in two search spaces on the ViT-Bench-101 [41] benchmark dataset. We tested ER-Score by incorporating it into the Auto-Prox search framework [41].

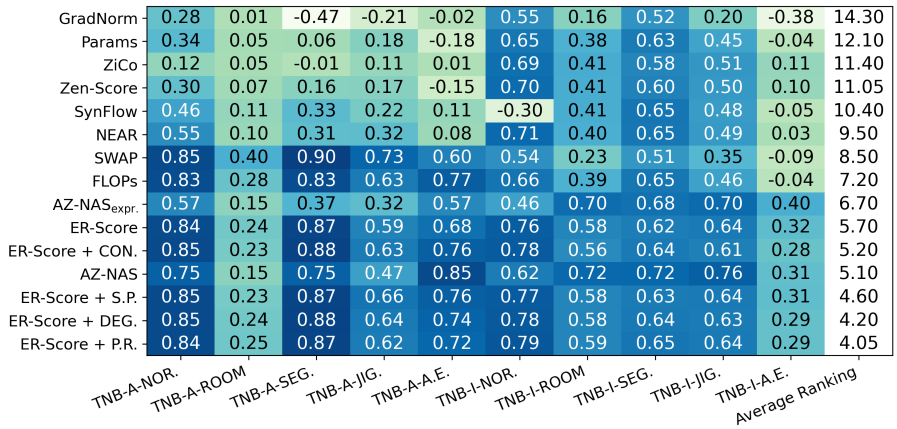


Fig. 6: Spearman's rank correlation coefficients between ZCPs and networks' ground-truth performance for 10 existing metrics and our methods on *downstream transfer tasks*. The rows are sorted based on average ranks of independent experiments for each metric. VKDNW [39] is not included as it is not applicable.

Table 2: Spearman's rank correlation coefficients (%) on ViT-Bench-101 [41]. Our method achieves the highest ranking correlation with distillation accuracy.

Search Space	Method	C100	Flowers	Chaoyang
AutoFormer [5]	AZ-NAS [21]	59.28	33.64	9.58
	TF-TAS [46]	50.84	83.28	38.52
	Auto-Prox [41]	63.87	83.65	41.76
	Ours	**95.36**	**89.17**	**53.58**
PiT [46]	AZ-NAS [21]	21.10	26.15	37.47
	TF-TAS [46]	82.20	82.91	52.92
	Auto-Prox [41]	95.12	92.94	55.09
	Ours	**95.90**	**96.33**	**76.14**

Overall, our method demonstrates strong and stable predictive performance, achieving state-of-the-art or competitive results against existing ZCPs across a wide range of tasks. Specifically, as shown in Figure 5, on image classification benchmarks (NB201, NB301, and TNB-A/I), ER-Score and its topology-weighted variants achieve the highest Spearman's rank correlation coefficients among all compared methods. Beyond classification, Figure 6 further confirms its effectiveness on downstream transfer tasks within the TNB-A/I benchmarks, demonstrating its generalization capability across diverse task types. Among the proposed topology-weighted variants, *ER-Score + DEG.* and *ER-Score + P.R.* achieve the highest ranking consistency, verifying the effectiveness of incorporating structural priors into ER-Score.

Furthermore, as shown in Table 2, we extend the evaluation to the AutoFormer [5] and PiT [46] search spaces on ViT-Bench-101, where 'Ours' denotes the integration of ER-Score into the Auto-Prox search framework. The results demonstrate that ER-Score also exhibits strong applicability beyond CNN architectures, achieving the highest ranking correlation with distillation accuracy in both Transformer-based search spaces. Detailed searched formulas can be found in the supplementary material.

Elite Architecture Discrimination Analysis To further investigate the discriminative capability of TER-Score in identifying high-performing architectures, we sample 10^5 architectures from the DARTS search space in NB301 and report the normalized discounted cumulative gain (nDCG) [3] over the top $\{1000, 500, 100,\ 50, 10\}$ ranked architectures, as shown in Figure 7. The nDCG is defined as:

$$\mathrm{nDCG}_P = \frac{1}{Z} \sum_{j=1}^{P} \frac{2^{\mathrm{acc}_{k_j}} - 1}{\log_2(1+j)}, \tag{13}$$

where $P \in \mathbb{N}$ denotes the number of top-ranked networks considered, and Z is a normalization factor corresponding to the ideal discounted cumulative gain, ensuring $\mathrm{nDCG}_P = 1$ for a perfect ranking.

The results indicate that although the *ER-Score + P.R.* variant shows relatively weaker performance in the global evaluation of NB301, it exhibits notably stronger discrimination capability among elite architectures, demonstrating its potential in fine-grained architecture ranking. We then **adopt the ER-Score + P.R. variant as the TER-Score for the following experiments** due to its potential to distinguish elite architectures.

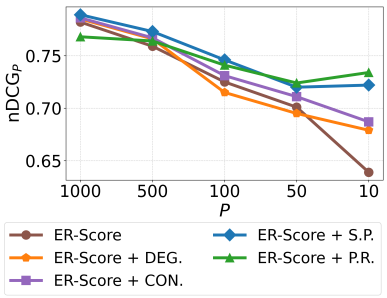


Fig. 7: The nDCG$_P$ of TER-Score on NB301 [45].

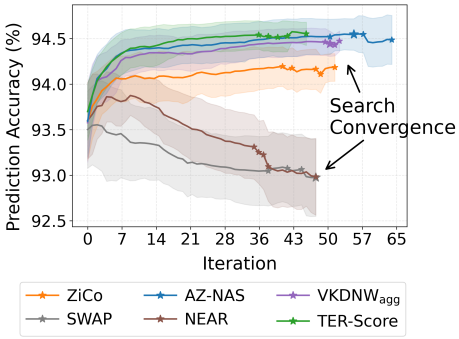


Fig. 8: Evolutionary search trajectories on NB301 [45].

4.2 Results on Evolutionary Search

To evaluate the effectiveness of the TER-Score in a larger search space, we conducted evolutionary search experiments using the NB301 surrogate model [45] on the DARTS [25] search space. We performed five evolutionary search experiments with different random seeds to ensure statistical robustness and mitigate the impact of randomness in the optimization process.

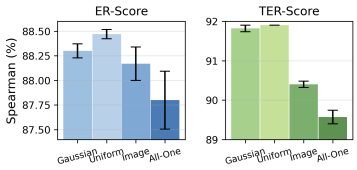
In Figure 8, 'Prediction Accuracy' indicates the prediction results of the surrogate model. ★ marks the termination of each individual search. Larger stars denote the simultaneous termination of two searches. The solid curve shows the mean value, and the shaded area represents the standard deviation across different runs, providing a comprehensive view of the search dynamics and stability.

The results show that the TER-Score **consistently surpasses the existing ZCPs throughout the entire search process**, achieving the highest final accuracy with the lowest performance variance. These results highlight the efficiency and robustness of the TER-Score as a reliable guidance signal for evolutionary architecture optimization. In contrast, some ZCPs experience a performance drop during the search process. This performance degradation may be attributed to their inherent preference for architectures with a large number of parameters, which is suboptimal for the small-scale C10 dataset, whereas our method maintains stable guidance by focusing on training dynamics stability rather than parameter count.

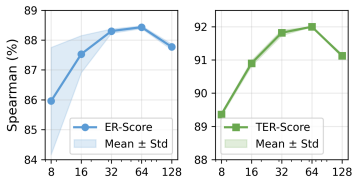
Table 3: Performance comparison of architectures discovered by our method and prior approaches on the C10 (left) and ImageNet-1k (right) datasets.

Method	Test Error (%)	Params (MB)	GPU Days
NASNet-A [49]	2.65 / 26.0	3.3 / 5.3	2000
DARTS [25]	2.76 / 26.7	3.3 / 4.7	4
SNAS [42]	2.85 / 27.3	2.8 / 4.3	1.5
GDAS [11]	2.82 / 27.5	2.5 / 4.4	0.17
PINAT [26]	2.54 / 24.9	3.6 / 5.2	0.3
PRE-NAS [40]	2.49 / 24.0	4.5 / 6.2	0.6
SWAP-NAS [31]	2.48 / 24.0	4.3 / 5.8	0.004
AZ-NAS [21]	2.79 / 23.7	4.7 / 6.4	0.06
NEAR [17]	3.27 / 23.8	4.6 / 6.4	0.08
$\text{VKDNW}_{\text{agg}}$ [39]	2.75 / 23.8	4.7 / 6.4	0.06
TER-Score	**2.41** / **23.55**	4.0 / 6.1	0.09

Table 4: Ablation study on mini-batch input. Results are derived from 3 independent runs.



(a) Distribution



(b) Resolution

4.3 Results of Full-Training on the DARTS Search Space

We trained the architectures discovered by TER-Score through evolutionary search and compared their performance with SOTA NAS methods, including multi-shot [49], one-shot [11,25,42], predictor-based [26,40], and training-free [17, 21, 31, 39] methods. For AZ-NAS [21], NEAR [17], and $\text{VKDNW}_{\text{agg}}$ [39], we trained the architectures discovered by the same framework, as the original papers did not report the corresponding results.

As shown in Table 3, the TER-Score achieves the **lowest test error of 2.41%** on the C10 dataset, surpassing all training-free and predictor-based methods with a low computational cost. On the more challenging ImageNet-1k dataset, TER-Score achieves the **lowest test error of 23.55%**, demonstrating strong generalization capability in large-scale vision tasks. These results collectively validate the superiority of our method in terms of both efficiency and accuracy across datasets of varying scales.

4.4 Ablation Study

Table 4 examines the impact of mini-batch construction on the effectiveness of our proposed scores.

Distribution. For input distribution, Gaussian and uniform random inputs yield the strongest correlations for both ER-Score and TER-Score, whereas real images are weaker, and all-one input degrades most clearly. This indicates that the method benefits from diverse, unbiased perturbations rather than semantic image content, which aligns with our training-dynamics perspective that emphasizes structural stability over data-specific statistics.

Resolution. For input resolution, performance improves from low resolutions and peaks at medium settings before dropping at high resolutions, suggesting a balance between feature diversity and redundant local patterns. This trend implies that moderately downsampled inputs preserve sufficient structural information while reducing computational overhead.

Overall, random inputs with moderate resolution provide the most reliable and efficient evaluation setting for practical deployment.

5 Limitations

Despite the strong empirical performance, our method has several limitations that should be acknowledged. First, the proposed topology-weighting strategies are heuristic, and providing a rigorous theoretical justification for the choice of specific graph-theoretic metrics on neural DAGs remains inherently challenging. This limitation is partially offset by the strong empirical results across multiple benchmarks, where the topology-weighted variants consistently outperform their unweighted counterparts. Second, the TER-Score is designed for CNNs and does not naturally generalize to Transformers, which restricts its applicability in modern vision tasks. However, the ER-Score itself is architecture-agnostic and can be directly applied to Transformers, as demonstrated in Table 2 on ViT-Bench-101. Future work could extend topology-weighting to Transformers via attention-based or token-level structural priors.

6 Conclusion

In this work, we study zero-cost architecture evaluation from the perspective of training dynamics and propose a principled framework for training-free NAS. We introduce the Effective Rank Score (ER-Score), which characterizes the stability of optimization behavior as a core criterion for architecture comparison beyond static architectural signals. We further incorporate topology-aware weighting to account for structural heterogeneity, yielding the topology-weighted TER-Score, an interpretable and data-independent evaluation of architectural quality.

Extensive experiments across multiple NAS benchmarks show that our method yields consistent architecture rankings under constrained computational budgets. Beyond empirical results, our work highlights the role of training dynamics and architectural structure as complementary factors in zero-cost evaluation and suggests a direction for developing more principled training-free NAS proxies.

7 Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 62376163 and 62306180; and in part by the Shenzhen Science and Technology Program under Grants JCYJ20250604181716022 and JCYJ20250604181503004.

References

1. Abdelfattah, M.S., Mehrotra, A., Dudziak, Ł., Lane, N.D.: Zero-Cost Proxies for Lightweight NAS. In: International Conference on Learning Representations (ICLR) (2021)
2. Baker, B., Gupta, O., Naik, N., Raskar, R.: Designing Neural Network Architectures using Reinforcement Learning. In: International Conference on Learning Representations (ICLR) (2017)
3. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to Rank using Gradient Descent. In: International Conference on Machine Learning (ICML). pp. 89–96 (2005)
4. Cavagnero, N., Robbiano, L., Caputo, B., Averta, G.: FreeREA: Training-Free Evolution-based Architecture Search. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1493–1502 (2023)
5. Chen, M., Peng, H., Fu, J., Ling, H.: AutoFormer: Searching Transformers for Visual Recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12270–12280 (2021)
6. Chen, W., Gong, X., Wang, Z.: Neural Architecture Search on ImageNet in Four GPU Hours: A Theoretically Inspired Perspective. In: International Conference on Learning Representations (ICLR) (2021)
7. Chrabaszcz, P., Loshchilov, I., Hutter, F.: A Downsampled Variant of ImageNet as an Alternative to the CIFAR datasets. arXiv preprint arXiv:1707.08819 (2017)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. IEEE (2009)
9. Dijkstra, E.W.: A Note on Two Problems in Connexion with Graphs. Numerische Mathematik **1**, 269–271 (1959)
10. Dong, P., Li, L., Tang, Z., Liu, X., Wei, Z., Wang, Q., Chu, X.: ParZC: Parametric Zero-Cost Proxies for Efficient NAS. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 16327–16335 (2025)
11. Dong, X., Yang, Y.: Searching for a Robust Neural Architecture in Four GPU Hours. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1761–1770 (2019)
12. Dong, X., Yang, Y.: NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search. In: International Conference on Learning Representations (ICLR) (2020)
13. Du, S.S., Zhai, X., Poczos, B., Singh, A.: Gradient descent provably optimizes over-parameterized neural networks. In: International Conference on Learning Representations (ICLR) (2019)
14. Duan, Y., Chen, X., Xu, H., Chen, Z., Liang, X., Zhang, T., Li, Z.: TransNAS-Bench-101: Improving Transferability and Generalizability of Cross-Task Neural Architecture Search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2021). `https://doi.org/10.1109/cvpr46437.2021.00521`
15. Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics pp. 1189–1232 (2001)
16. Guo, Z., Zhang, X., Mu, H., Heng, W., Liu, Z., Wei, Y., Sun, J.: Single Path One-Shot Neural Architecture Search with Uniform Sampling. In: European Conference on Computer Vision (ECCV). pp. 544–560. Springer (2020)

17. Husistein, R.T., Reiher, M., Eckhoff, M.: NEAR: A Training-Free Pre-Estimator of Machine Learning Model Performance. In: International Conference on Learning Representations (ICLR) (2025)
18. Ji, H., Feng, Y., Fan, J., Sun, Y.: CARL: Causality-guided Architecture Representation Learning for an Interpretable Performance Predictor. arXiv preprint arXiv:2506.04001 (2025)
19. Jiang, T., Wang, H., Bie, R.: MeCo: Zero-Shot NAS with One Data and Single Forward Pass via Minimum Eigenvalue of Correlation. Advances in Neural Information Processing Systems (NeurIPS) **36**, 61020–61047 (2023)
20. Krizhevsky, A., Hinton, G., et al.: Learning Multiple Layers of Features from Tiny Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Toronto, ON, Canada (2009)
21. Lee, J., Ham, B.: AZ-NAS: Assembling Zero-Cost Proxies for Network Architecture Search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5893–5903 (2024)
22. Li, G., Hoang, D., Bhardwaj, K., Lin, M., Wang, Z., Marculescu, R.: Zero-Shot Neural Architecture Search: Challenges, Solutions, and Opportunities. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
23. Li, G., Yang, Y., Bhardwaj, K., Marculescu, R.: ZiCo: Zero-shot NAS via Inverse Coefficient of Variation on Gradients. In: International Conference on Learning Representations (ICLR) (2023)
24. Lin, M., Wang, P., Sun, Z., Chen, H., Sun, X., Qian, Q., Li, H., Jin, R.: Zen-NAS: A Zero-Shot NAS for High-Performance Image Recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 347–356 (2021)
25. Liu, H., Simonyan, K., Yang, Y.: DARTS: Differentiable Architecture Search. In: International Conference on Learning Representations (ICLR) (2019)
26. Lu, S., Hu, Y., Wang, P., Han, Y., Tan, J., Li, J., Yang, S., Liu, J.: PINAT: A Permutation INvariance Augmented Transformer for NAS Predictor. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 8957–8965 (2023)
27. Lu, S., Hu, Y., Yang, L., Sun, Z., Mei, J., Tan, J., Song, C.: PA&DA: Jointly Sampling Path and Data for Consistent NAS. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11940–11949 (2023)
28. Nilsback, M.E., Zisserman, A.: Automated Flower Classification over a Large Number of Classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
29. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Tech. rep., Stanford infolab (1999)
30. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An Imperative Style, High-performance Deep Learning Library. Advances in Neural Information Processing Systems (NeurIPS) **32** (2019)
31. Peng, Y., Song, A., Fayek, H.M., Ciesielski, V., Chang, X.: SWAP-NAS: Sample-Wise Activation Patterns for Ultra-fast NAS. In: International Conference on Learning Representations (ICLR) (2024)
32. Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., et al.: MobileNetV4: Universal Models for the Mobile Ecosystem. In: European Conference on Computer Vision (ECCV). pp. 78–96. Springer (2024)

33. Roy, O., Vetterli, M.: The effective rank: A measure of effective dimensionality. In: 15th European signal processing conference. pp. 606–610. IEEE (2007)
34. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4510–4520 (2018)
35. Shu, Y., Dai, Z., Wu, Z., Low, B.K.H.: Unifying and Boosting Gradient-Based Training-Free Neural Architecture Search. Advances in Neural Information Processing Systems (NeurIPS) **35**, 33001–33015 (2022)
36. Siddiqui, S., Kyrkou, C., Theocharides, T.: Operation and Topology Aware Fast Differentiable Architecture Search. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9666–9673. IEEE (2021)
37. Spearman, C.: The Proof and Measurement of Association between Two Things. The American Journal of Psychology **15**(1), 72–101 (1904)
38. Tanaka, H., Kunin, D., Yamins, D.L., Ganguli, S.: Pruning neural networks without any data by iteratively conserving synaptic flow. Advances in neural information processing systems (NeurIPS) **33**, 6377–6389 (2020)
39. Tybl, O., Neumann, L.: Training-Free Neural Architecture Search through Variance of Knowledge of Deep Network Weights. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14881–14890 (2025)
40. Wang, H., Ge, C., Chen, H., Sun, X.: PreNAS: Preferred One-Shot Learning Towards Efficient Neural Architecture Search. In: International Conference on Machine Learning (ICML). pp. 35642–35654. PMLR (2023)
41. Wei, Z., Dong, P., Hui, Z., Li, A., Li, L., Lu, M., Pan, H., Li, D.: Auto-Prox: Training-Free Vision Transformer Architecture Search via Automatic Proxy Discovery. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 15814–15822 (2024)
42. Xie, S., Zheng, H., Liu, C., Lin, L.: SNAS: Stochastic Neural Architecture Search. In: International Conference on Learning Representations (ICLR) (2019)
43. Yang, Y., Zhang, X., Zhu, Q., Chen, W., Wong, K.C., Lin, Q.: CAGAN: Constrained Neural Architecture Search for GANs. Knowledge-Based Systems **302**, 112277 (2024)
44. Yang, Y., Zhu, Q., Luo, J., Wong, K.C., Lin, Q., Li, J.: TRNAS: A Training-Free Robust Neural Architecture Search. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2336–2345 (2025)
45. Zela, A., Siems, J., Zimmer, L., Lukasik, J., Keuper, M., Hutter, F.: Surrogate NAS benchmarks: Going beyond the limited search spaces of tabular NAS benchmarks. In: International Conference on Learning Representations (ICLR) (2022)
46. Zhou, Q., Sheng, K., Zheng, X., Li, K., Sun, X., Tian, Y., Chen, J., Ji, R.: Training-Free Transformer Architecture Search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10894–10903 (2022)
47. Zhu, C., Chen, W., Peng, T., Wang, Y., Jin, M.: Hard Sample Aware Noise Robust Learning for Histopathology Image Classification. IEEE Transactions on Medical Imaging **41**(4), 881–894 (2021)
48. Zoph, B., Le, Q.V.: Neural Architecture Search with Reinforcement Learning. In: International Conference on Learning Representations (ICLR) (2017)
49. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning Transferable Architectures for Scalable Image Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8697–8710 (2018)