Info-Driven Zero-Cost Proxy: Rethinking Vision Transformer Architecture Evaluation via Information Quantification

Yue Yang[1,†], **Jiacheng Wang**[1,†], **Zhenkai Yang**[1], **Menglan Hu**[1], **Gaoyang Liu**[1], **Bo Xu**[1,2], **Tianyue Zheng**[3] and **Kai Peng**[1,*]
[1]Huazhong University of Science and Technology
[2]Hubei ChuTianYun Co.,Ltd.
[3]Southern University of Science and Technology
{yangyue, wangjiacheng, zk_yang, humenglan, liugaoyang, xubosite, pkhust}@hust.edu.cn, zhengty@sustech.edu.cn

Abstract

Neural Architecture Search (NAS) automates the design of Vision Transformer (ViT) architectures. However, the high computational cost of training-based methods has made zero-cost proxies a key research direction. While existing proxies can estimate model potential, they fail to capture critical information transmission characteristics due to the black-box nature of neural networks. Consequently, they suffer from high computational latency and weak correlation with ground-truth performance. Although information transmission determines performance, it has not been effectively quantified. This limitation remains the bottleneck for current methods. In this paper, we propose Info-NAS, a zero-cost proxy based on architectural information. This method achieves the first structured quantification of information transmission in ViT. Specifically, it evaluates architectures using three core components: global information volume, local information gradient, and global consistency. Info-NAS derives proxy scores solely from architectural parameters, requiring no forward or backward propagation. Consequently, the calculation and search processes are highly efficient. Moreover, we construct the ViT-Info-Bench dataset for algorithm evaluation. Experimental results demonstrate that Info-NAS significantly reduces search overhead while achieving superior ranking accuracy. With a single evaluation requiring only 1.8 ms, Info-NAS outperforms existing methods in both efficiency and performance. Our code is available at https://github.com/Shallow-W/Info-NAS.

1 Introduction

In recent years, Vision Transformer (ViT)[Dosovitskiy *et al.*, 2021; Liu *et al.*, 2021] has demonstrated exceptional modeling capabilities in computer vision tasks, gradually becoming a mainstream architecture in the field. However, the performance of ViT relies heavily on complex architectural designs [Touvron *et al.*, 2021b; Zhai *et al.*, 2022; Alabdulmohsin *et al.*, 2023], including the selection of hyperparameters such as embedding dimension, number of heads, MLP expansion ratio, and network depth. The search space formed by these design dimensions is vast, where even subtle structural differences can lead to drastic fluctuations in model performance. Traditional manual design approaches are not only time-consuming and labor-intensive but also rely heavily on expert experience, making it difficult to systematically explore optimal structures [Ren *et al.*, 2021; Gong and Wang, 2022].

To automate the design of efficient ViT architectures, traditional Neural Architecture Search (NAS) [Elsken *et al.*, 2019; Wei *et al.*, 2024b] emerged as an early solution. However, this approach requires training every candidate model, consuming massive computational resources. To reduce costs, researchers proposed supernet [Guo *et al.*, 2020; Chen *et al.*, 2021b] methods, which involve training a single colossal network and subsequently selecting different sub-structures from

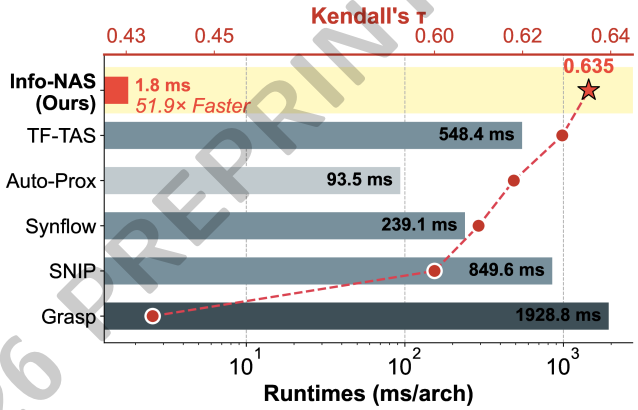


Figure 1: Comparison of zero-cost proxies on the CIFAR-100 dataset from ViT-Info-Bench. We compare Info-NAS with existing zero-cost proxies in terms of runtime and Kendall's τ. Info-NAS requires only 1.8 ms per architecture, which is 51.9× faster than Auto-Prox [Wei *et al.*, 2024a], while achieving the highest correlation. This demonstrates that Info-NAS outperforms other zero-cost proxies in both performance and efficiency.

*Corresponding author.

it. Although faster than traditional methods, training a supernet still requires hundreds of GPU hours. Consequently, Training-free NAS [Abdelfattah *et al.*, 2021; Chen *et al.*, 2021c; Mellor *et al.*, 2021] has become a research hotspot. The core of this approach lies in designing a proxy metric capable of predicting the final performance of an architecture at Zero-Cost, enabling the rapid evaluation of untrained models and reducing search costs.

However, the search efficiency of mainstream zero-cost proxy [Lee *et al.*, 2019; Tanaka *et al.*, 2020; Ha *et al.*, 2023; Jiang *et al.*, 2023], including ViT-specific approaches such as TF-TAS [Zhou *et al.*, 2024] and Auto-Prox [Wei *et al.*, 2024a], remains suboptimal. Since these methods rely on forward or backward propagation, they incur non-negligible computational overhead when applied to large-scale search spaces. Furthermore, these methods often overlook the structural characteristics of ViT and rely on single metrics, thereby failing to provide a comprehensive evaluation of the architectures. Finally, although some studies have attempted to establish benchmark datasets [Ho *et al.*, 2024; Wei *et al.*, 2024a] for ViT architecture search, they remain insufficient in terms of dataset types and the number of architectures. This limitation directly hinders the rigorous evaluation of zero-cost proxy performance.

To address these issues, this paper proposes Info-NAS, a zero-cost proxy method based on information quantification that deeply models the unique structural characteristics of ViT by quantifying their information features. This represents the first structural modeling and quantification of the information processing efficiency of ViT architectures, providing a new physical perspective for understanding the intrinsic quality of neural network architectures. Specifically, we define the concept of Layer-wise Modulated Information and evaluate the potential of ViT architectures using three core components: (1) Global Information Volume ($\mathcal{V}$): Measures the inherent information processing capability of network layers. (2) Local Information Gradient ($\mathcal{G}$): Evaluates the smoothness of information flow between layers. (3) Global Consistency ($\mathcal{C}$): Analyzes the overall consistency of the architecture. Finally, we fuse the information features from these three dimensions to obtain the final proxy score. Additionally, we constructed the ViT-Info-Bench dataset, which contains diverse ViT architectures and their ground-truth performance on CIFAR-10, CIFAR-100, Mini-ImageNet, and Tiny-ImageNet, providing a reliable benchmark for evaluating zero-cost proxy research on ViT.

Our core contributions can be summarized as follows:

- We propose Info-NAS, a zero-cost proxy that achieves the first structured quantification of information transmission in ViT. This method evaluates the architecture through three core components: Global Information Volume, Local Information Gradient, and Global Consistency, without requiring any forward or backward propagation.
- We construct a benchmark dataset based on the AutoFormer [Chen *et al.*, 2021b] search space, which includes various ViT architectures and their performance across multiple datasets. This establishes an evaluation platform for training-free methods and provides data support for the correlation analysis of zero-cost proxies.
- Through extensive experiments, we comprehensively validate the superiority of Info-NAS across multiple benchmark datasets and search spaces. The results demonstrate that Info-NAS outperforms advanced zero-cost proxies in terms of the correlation between proxy scores and ground-truth accuracy. More importantly, as shown in Figure 1, while maintaining high performance, Info-NAS significantly reduces the computational overhead of proxy scoring and the overall search time, achieving search efficiency far surpassing existing training-free algorithms.

2 Related Works

2.1 Neural Architecture Search for ViT

Early ViT NAS adopted the One-Shot [Guo *et al.*, 2020; Peng *et al.*, 2020] paradigm from CNN. The core strategy is to train a supernet that contains all candidate sub-networks. AutoFormer [Chen *et al.*, 2021b] pioneered this direction by introducing a weight entanglement mechanism. It achieved superior performance by evaluating sub-networks sampled from the trained supernet. Subsequently, ViTAS [Su *et al.*, 2022] proposed cyclic weight sharing and identity transformation mechanisms. Independent class tokens were assigned to different patch sizes to reduce weight interference between sub-networks. Later methods focused on more complex search spaces and optimization strategies. For example, GLiT [Chen *et al.*, 2021a] introduced convolutional operations to search for hybrid architectures. NASViT [Gong and Wang, 2022] addressed gradient conflicts during supernet training and used hardware constraints to guide the search. Although these methods automate architecture design, supernet training remains extremely expensive. The high computational cost hinders the application of NAS in scenarios requiring rapid iteration.

2.2 Zero-Cost Proxy

Training-free architecture search has emerged as a promising direction to mitigate the computational costs of supernet training. These methods utilize a zero-cost proxy to predict final performance at initialization, bypassing expensive training processes. Early proxies [Lee *et al.*, 2019; Tanaka *et al.*, 2020; Abdelfattah *et al.*, 2021] primarily targeted CNN and relied on weight magnitudes or gradient norms. Recently, proxies tailored for ViT have been developed. TF-TAS [Zhou *et al.*, 2024], the pioneering training-free method for ViT, assesses architectural potential by measuring synaptic diversity in MSA and synaptic saliency in MLP. Auto-Prox [Wei *et al.*, 2024a] introduced an automated discovery framework to enhance robustness using a Joint Correlation Metric (JCM). AZ-NAS [Lee and Ham, 2024] improved ranking consistency by integrating metrics for expressivity, progressivity, trainability, and complexity. However, existing ViT proxies typically require a forward or backward pass, preventing them from being truly zero-cost. In this paper, we model the architectural information flow based on ViT components and propose a simpler, effective proxy. This

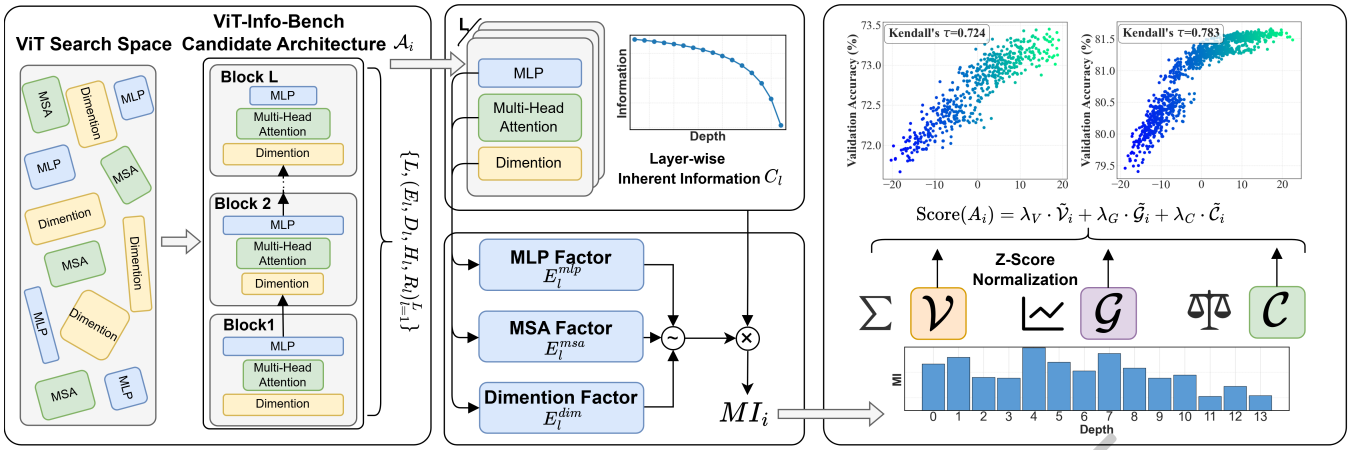


Figure 2: The overall framework of Info-NAS. Given a candidate architecture A_i sampled from the ViT search space, we first model the Layer-wise Inherent Information (C_l) based on network depth and calculate the modulation factors for Dimension, MSA, and MLP. These components are integrated to derive the Layer-wise Modulated Information (MI_l). Subsequently, we employ three components to evaluate the architecture: $\mathcal{V}$, $\mathcal{G}$, $\mathcal{C}$. The final proxy score is obtained by a weighted summation of these metrics after Z-score normalization, enabling efficient performance prediction without training.

approach drastically reduces evaluation time while significantly improving NAS performance.

3 Methodology

Figure 2 illustrates the overall framework of the proposed information-driven zero-cost proxy method. First, we introduce the concepts of Layer-wise Inherent Information, Information Modulation Coefficient, and Layer-wise Modulated Information. Next, we define three components by quantifying the connection between architectural parameters and information efficiency. Finally, we normalize these metrics to formulate the comprehensive proxy score.

3.1 Layer-wise Information Modeling

Within a discrete ViT architecture search space $\mathcal{S}$, any candidate architecture $\mathcal{A}$ is uniquely determined by the following tuple of hyperparameters: $\mathcal{A} = \{L, (E_l, D_l, H_l, R_l)_{l=1}^{L}\}$ where L denotes the total number of layers in the Transformer encoder. For the l-th layer, E_l represents the embedding dimension; D_l indicates the dimension of the Q-K-V matrices; H_l signifies the number of attention heads; and R_l denotes the expansion ratio of MLP.

Layer-wise Inherent Information

In deep neural networks, information transmission often decreases layer by layer after multiple non-linear transformations [He *et al.*, 2016; Liu *et al.*, 2022]. This information dissipation extends to ViT [Noci *et al.*, 2022; Dong *et al.*, 2021; Zhou *et al.*, 2021], where deep features become homogenized, implying severe information loss during transmission. Motivated by this, we introduce Layer-wise Inherent Information. Treating the network as a transmission channel with inherent loss, we employ a depth-based exponential decay model to simulate natural information attenuation in an idealized state. Specifically, for the l-th layer of architecture $\mathcal{A}$, it is defined as:

$$C_l = 2^{\alpha - \frac{\beta}{\gamma - l}}, \tag{1}$$

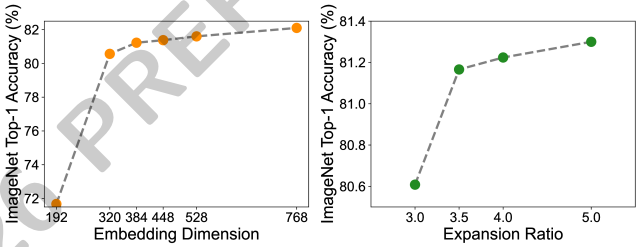


Figure 3: Impact of architectural hyperparameters on ImageNet Top-1 accuracy. The plots visualize the performance trends when varying the Embedding Dimension (Left) and Expansion Ratio (Right).

where l represents the layer depth, and α, β, γ are hyperparameters controlling the inherent information decay curve. We set $\alpha = 12$ and $\beta = 2$ through experiments. Given the maximum search depth of 16, we assign $\gamma = 17$, where this slight margin ensures γ functions as a valid asymptotic boundary.

Information Modulation Coefficient

While C_l serves as a theoretical prior, the effective information transmission is governed by specific architectural configurations that actively modulate the information flow. Consequently, we introduce the Information Modulation Coefficient η to quantify this effect for the l-th layer. Considering that a ViT layer comprises MSA and MLP blocks operating within a defined embedding dimension, η is designed to jointly evaluate the efficacy of these structural components:

$$\eta_l = (E_l^{dim})^{w_{dim}} \cdot (E_l^{msa})^{w_{msa}} \cdot (E_l^{mlp})^{w_{mlp}}, \tag{2}$$

where E_l^{dim}, E_l^{msa}, and E_l^{mlp} represent the embedding dimension factor, the MSA factor, and the MLP factor of the l-th layer, respectively; w_{dim}, w_{msa}, and w_{mlp} are hyperparameters used to regulate the sensitivity of each factor in the overall score.

Dimension Factor. Serving as the carrier of information flow, the embedding dimension E_l defines the width of the inter-layer channel and determines the capacity upper bound of the feature state space. To quantify its impact on processing efficiency, we analyze how model performance responds to varying dimensions. As shown in Figure 3, widening the channel significantly enhances feature representation in the low-dimension regime. However, performance quickly saturates with further expansion and approximates a logarithmic trend. Consequently, we quantify the gain of E_l logarithmically and define the factor as:

$$E_l^{dim} = \log E_l. \tag{3}$$

MSA Factor. To quantify the efficacy of the MSA module, we model it through two dimensions: Feature Transformation Scale and Multi-head Subspace Diversity. First, we analyze the feature transformation scale. As the search space allows for decoupled embedding dimensions E_l and Q-K-V dimensions D_l, a linear projection is required within the MSA module. Specifically, the mapping from the embedding space $\mathbb{R}^{E_l}$ to the Q-K-V subspaces $\mathbb{R}^{D_l}$ is performed by a weight matrix $\mathbf{W}_l \in \mathbb{R}^{E_l \times D_l}$. The scale of this transformation is determined by the product of the input and output dimensions, which quantifies both computational complexity and parameter volume. Therefore, the feature transformation scale $\mathcal{S}_l$ of the l-th layer is defined as:

$$\mathcal{S}_l = E_l \cdot D_l. \tag{4}$$

Theoretically, a restricted scale may cause an information bottleneck, leading to semantic loss. Conversely, an excessive scale introduces redundancy with minimal gain. To balance representation capacity and parameter efficiency, we propose an adaptability metric $P(\mathcal{S}_l)$ utilizing the Sigmoid function.

$$P(\mathcal{S}_l) = \frac{1}{1 + \exp(-\sigma_1 \cdot (\mathcal{S}_l - \tau_1))}, \tag{5}$$

where σ_1 and τ_1 are tunable hyperparameters that control the steepness and threshold of the function, respectively.

Next is Multi-head Subspace Diversity. MSA projects inputs into H_l parallel subspaces to capture multi-perspective feature representations. While increasing H_l initially enriches semantics, the attentional foci of distinct heads inevitably overlap as the head count grows. This redundancy leads to diminishing marginal returns in non-redundant information gain. To quantify this saturation effect, we define multi-head subspace diversity as:

$$F(H_l) = 1 + \sigma_2 \cdot \log(H_l + \tau_2), \tag{6}$$

where σ_2 and τ_2 are hyperparameters that adjust the sensitivity and offset of the logarithmic function, respectively.

Synthesizing these dimensions, we define the MSA factor E_l^{msa} for the l-th layer as the product of the feature transformation scale and multi-head subspace diversity:

$$E_l^{msa} = P(\mathcal{S}_l(E_l, D_l)) \cdot F(H_l). \tag{7}$$

MLP Factor. The MLP performs nonlinear transformations on the information flow across channel dimensions. The expansion ratio R_l determines the capacity of the hidden feature space, directly reflecting the information reconstruction capability of that layer. To quantify the specific impact of R_l on performance, we conducted experiments. As shown in Figure 3, within the commonly used search range of $[3, 5]$, model accuracy demonstrates a monotonic ascent with increasing R_l. Therefore, we directly adopt R_l as the MLP factor:

$$E_l^{mlp} = R_l. \tag{8}$$

Layer-wise Modulated Information

Finally, we define the information of the l-th layer processed by architectural modulation as the Layer-wise Modulated Information. This metric is formed by coupling the layer-wise inherent information with the information modulation coefficient, defined as:

$$MI_l = C_l \cdot \eta_l. \tag{9}$$

3.2 Zero Proxy of Info-NAS

Next, we further analyze and process the Layer-wise Modulated Information and specifically design three core evaluation components.

Global Information Volume. In constructing the Info-NAS comprehensive proxy metric, the primary task is to quantify the architecture's capacity for information flow carriage and processing at the global scale. Since each network layer performs nonlinear mapping on input features, its contribution to the global representation is cumulative. Therefore, we define the Global Information Volume $\mathcal{V}$ by summing the modulated information of all layers.

$$\mathcal{V} = \sum_{i=1}^{L} MI_i. \tag{10}$$

Local Information Gradient. In training-free NAS research for ViT [Wei *et al.*, 2024a; Zhou *et al.*, 2024], gradient distribution and magnitude are considered direct indicators of model diversity. These metrics reflect the network's sensitivity to input features and its processing potential. Inspired by this, we introduce the Local Information Gradient component to capture feature transformation activity from the perspective of local dynamic evolution. To quantify the dynamic variations in inter-layer information flow, we define this metric as the cumulative sum of the change intensity of weighted information between adjacent layers:

$$\mathcal{G} = \sum_{i=2}^{L-1} |C_{i+1} \cdot MI_{i+1} - C_{i-1} \cdot MI_{i-1}|. \tag{11}$$

Global Consistency. To ensure macro structural robustness and prevent pathological architectures, we introduce the Global Consistency metric. Designed to enforce global equilibrium, it is defined as the sum of pairwise differences in weighted information between any two network layers:

$$\mathcal{C} = -\sum_{i=1}^{L} \sum_{j=i+1}^{L} |C_i \cdot MI_i - C_j \cdot MI_j|. \tag{12}$$

Method	CIFAR-10-T		CIFAR-100-T		Mini-ImageNet-T		Tiny-ImageNet-T		CIFAR-10-S		CIFAR-100-S	
	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman
Jacobian [Hosseini *et al.*, 2021]	0.195	0.285	0.236	0.345	0.244	0.363	0.241	0.355	0.292	0.426	0.248	0.371
Meco [Jiang *et al.*, 2023]	0.047	0.070	0.059	0.088	0.093	0.138	0.073	0.110	0.058	0.087	0.044	0.067
TE-NAS [Chen *et al.*, 2021c]	0.401	0.594	0.478	0.677	0.482	0.680	0.471	0.666	0.497	0.709	0.471	0.675
GraSP [Wang *et al.*, 2020]	0.292	0.438	0.385	0.555	0.382	0.550	0.355	0.512	0.470	0.666	0.436	0.626
Croze [Ha *et al.*, 2023]	0.485	0.709	0.608	0.815	0.615	0.824	0.591	0.804	0.640	0.850	0.609	0.818
SNIP [Lee *et al.*, 2019]	0.466	0.696	0.605	0.817	0.626	0.836	0.581	0.794	0.654	0.857	0.600	0.809
Synflow [Tanaka *et al.*, 2020]	0.508	0.734	**0.665**	**0.860**	0.710	0.896	0.649	0.848	0.659	0.864	0.610	0.818
TF-TAS [Zhou *et al.*, 2024]	0.552	0.776	0.591	0.801	0.711	0.896	0.646	0.846	0.650	0.857	0.629	0.835
Auto-Prox [Wei *et al.*, 2024a]	0.536	0.759	0.605	0.807	0.706	0.891	0.650	0.848	0.631	0.841	0.618	0.826
Ours	**0.566**	**0.783**	0.612	0.812	**0.724**	**0.902**	**0.679**	**0.868**	**0.666**	**0.866**	**0.635**	**0.840**

Table 1: Comparison of ranking correlations on the ViT-Info-Bench dataset. We report Kendall's τ and Spearman's ρ coefficients for both AutoFormer-Tiny (T) and AutoFormer-Small (S) search spaces across CIFAR-10, CIFAR-100, Mini-ImageNet, and Tiny-ImageNet datasets.

We formulate $\mathcal{C}$ as a structural penalty term. A value of $\mathcal{C}$ closer to zero implies a more stable distribution of information flow across the network depth, indicating superior global consistency.

Comprehensive Proxy Score. Based on the aforementioned research, we can directly compute the raw scores $\{\mathcal{V}_i, \mathcal{G}_i, \mathcal{C}_i\}$ for the three components using the architectural parameters. To eliminate dimensional differences between metrics of varying scales, we standardize each component's score using Z-scores. Finally, we obtain the comprehensive proxy score by linearly weighting and aggregating the three components:

$$\text{Score}(A_i) = \lambda_{\mathcal{V}} \cdot \tilde{\mathcal{V}}_i + \lambda_{\mathcal{G}} \cdot \tilde{\mathcal{G}}_i + \lambda_{\mathcal{C}} \cdot \tilde{\mathcal{C}}_i, \qquad (13)$$

where $\lambda_{\mathcal{V}}, \lambda_{\mathcal{G}}, \lambda_{\mathcal{C}}$ are adjustable weighting hyperparameters used to adjust the contribution weights among different components.

3.3 Training-free ViT NAS

We randomly sample 100,000 architectures from the constrained search space and compute the raw scores $\{\mathcal{V}_i, \mathcal{G}_i, \mathcal{C}_i\}$ for each architecture. Based on the composite proxy score derived from standardizing and weighting these three components, we select the network with the highest proxy score as the optimal architecture. Finally, we retrain it to obtain its test accuracy on the target dataset.

4 Experiments

In this section, we first outline the experimental setup, followed by an introduction to the benchmark datasets: ViT-Info-Bench, OoD-ViT-NAS, and ViT-Bench-101. Finally, we compare Info-NAS with existing methods and present a performance evaluation across multiple benchmarks.

4.1 Experimental Settings

We tuned the hyperparameters of Info-NAS on a small dataset containing 50 real architectures. To ensure fair evaluation, this dataset does not overlap with any subsequent benchmarks. The determined parameters were fixed for all experiments and were not fine-tuned for specific datasets. We evaluated our method on ViT-Info-Bench, ViT-Bench-101 [Wei *et al.*, 2024a], and OoD-ViT-NAS [Ho *et al.*, 2024]. These experiments cover different model complexities, with parameter counts ranging from 4-9M, 14-34M, and 42-75M. For evaluation, we select the architecture with the highest zero-cost proxy score and report its ground-truth accuracy. The details of these benchmark datasets are introduced below.

Method	Param (M)	Top-1 Acc (%)	Runtime (ms/arch)
CIFAR-100-T			
Meco [Jiang *et al.*, 2023]	9.43	70.58	1653.5
GraSP [Wang *et al.*, 2020]	5.68	69.57	1654.1
SNIP [Lee *et al.*, 2019]	9.80	70.72	813.6
Synflow [Tanaka *et al.*, 2020]	9.09	70.77	94.9
TF-TAS [Zhou *et al.*, 2024]	9.93	70.87	265.4
Auto-Prox [Wei *et al.*, 2024a]	9.76	70.99	88.6
Ours	9.98	**71.00**	**1.8**
CIFAR-100-S			
Meco [Jiang *et al.*, 2023]	17.86	74.36	1838.1
GraSP [Wang *et al.*, 2020]	21.35	74.92	1928.8
SNIP [Lee *et al.*, 2019]	31.52	75.13	849.6
Synflow [Tanaka *et al.*, 2020]	31.75	75.14	114.3
TF-TAS [Zhou *et al.*, 2024]	29.74	74.99	548.4
Auto-Prox [Wei *et al.*, 2024a]	29.74	74.99	93.5
Ours	31.94	**75.14**	**1.8**

Table 2: Evaluation of Top-1 accuracy and runtime efficiency on CIFAR-100-T and CIFAR-100-S. Info-NAS achieves the highest accuracy with lower computational latency compared to other zero-cost proxies.

ViT-Info-Bench. ViT-Info-Bench records the ground-truth accuracy of ViTs on four datasets: CIFAR-10, CIFAR-100, Mini-ImageNet, and Tiny-ImageNet. It is built upon the AutoFormer search space. We first trained the AutoFormer-T and AutoFormer-S supernets using unified settings for the optimizer, learning rate, and regularization to ensure parameter reliability. Subnets were then sampled to inherit pretrained weights directly from the supernet for accuracy evaluation. This benchmark includes 500 evaluated architectures per dataset, providing a basis for zero-cost proxy analysis.

ViT-Bench-101. ViT-Bench-101 provides both distillation and vanilla accuracies for ViT on small datasets, including CIFAR-100, Flowers, and Chaoyang, and the large-scale ImageNet-1K. It supports the evaluation of zero-cost proxies by analyzing score-accuracy correlations across different scenarios.

OoD-ViT-NAS. OoD-ViT-NAS is the first comprehensive benchmark for the OOD generalization of ViT in NAS. It

Method	CIFAR-100	Flowers	Chaoyang
GraSP [Wang *et al.*, 2020]	0.013	-0.007	-0.065
Synflow [Tanaka *et al.*, 2020]	0.529	0.821	0.380
TE-NAS [Chen *et al.*, 2021c]	0.433	0.738	0.394
NWOT [Mellor *et al.*, 2021]	0.631	0.821	0.384
TF-TAS [Zhou *et al.*, 2024]	0.508	0.833	0.381
Auto-Prox [Wei *et al.*, 2024a]	0.639	0.836	**0.411**
Ours	**0.769**	**0.846**	0.384

Table 3: Spearman correlation results on ViT-Bench-101 across different datasets.

Method	Kendall	Spearman	Pearson
OoD-ImageNet-T			
Meco [Jiang *et al.*, 2023]	0.328	0.483	0.504
TE-NAS [Chen *et al.*, 2021c]	0.218	0.334	0.403
Croze [Ha *et al.*, 2023]	0.270	0.402	0.409
SNIP [Lee *et al.*, 2019]	0.251	0.378	0.384
Synflow [Tanaka *et al.*, 2020]	0.404	0.575	0.410
TF-TAS [Zhou *et al.*, 2024]	0.437	0.621	0.654
Auto-Prox [Wei *et al.*, 2024a]	0.477	0.675	0.730
Ours	**0.655**	**0.844**	**0.858**
OoD-ImageNet-S			
Meco [Jiang *et al.*, 2023]	0.227	0.336	0.272
TE-NAS [Chen *et al.*, 2021c]	0.432	0.637	0.735
Croze [Ha *et al.*, 2023]	0.727	0.906	0.862
SNIP [Lee *et al.*, 2019]	0.720	0.904	0.874
Synflow [Tanaka *et al.*, 2020]	0.774	0.935	0.879
TF-TAS [Zhou *et al.*, 2024]	0.752	0.926	0.878
Auto-Prox [Wei *et al.*, 2024a]	0.742	0.922	0.867
Ours	**0.783**	**0.938**	**0.900**
OoD-ImageNet-B			
Meco [Jiang *et al.*, 2023]	0.124	0.185	0.181
TE-NAS [Chen *et al.*, 2021c]	0.098	0.150	0.161
Croze [Ha *et al.*, 2023]	0.150	0.250	0.229
SNIP [Lee *et al.*, 2019]	0.154	0.263	0.244
Synflow [Tanaka *et al.*, 2020]	0.088	0.173	0.131
TF-TAS [Zhou *et al.*, 2024]	0.060	0.118	0.109
Auto-Prox [Wei *et al.*, 2024a]	0.012	0.005	0.010
Ours	**0.175**	**0.294**	**0.304**

Table 4: Correlation analysis (Kendall, Spearman, Pearson) on the ImageNet dataset of the OoD-ViT-NAS benchmark. Info-NAS consistently maintains the highest correlation across all metrics.

comprises 3,000 architectures, with 1,000 sampled from each of the AutoFormer Tiny, Small, and Base search spaces. These architectures are evaluated on eight mainstream OOD datasets, such as ImageNet-C, providing core metrics including test accuracy and parameter counts.

4.2 Experimental Results On ViT-Info-Bench

We evaluated the rank correlation of various NAS methods on ViT-Info-Bench using Kendall's τ [Abdi, 2007] and Spearman's ρ [Stephanou and Varughese, 2021]. As shown in Table 1 and Figure 4, Info-NAS consistently outperforms competitors across most metrics. These results demonstrate that Info-NAS accurately identifies high-performance ViT architectures and exhibits strong robustness across diverse datasets. Further analysis reveals that methods such as Jacobian and Meco primarily focus on local features or gradient sensitivity. While effective for CNN, these approaches struggle in the ViT search space due to gradient noise and signal attenuation during initialization. In contrast, Info-NAS quantifies discrete architectural parameters as information transmission capabilities, directly evaluating performance through these parameters and deeply modeling the structural characteristics of ViT across the $\mathcal{V}$, $\mathcal{G}$, and $\mathcal{C}$ dimensions. This result not only validates the comprehensiveness and precision of Info-NAS in capturing ViT performance attributes but also underscores the significant potential and rationality of constructing efficient zero-cost proxies centered on information flow.

Models	Param (M)	Top-1 (%)	Type	GPU Days
Deit-Ti [Touvron *et al.*, 2021a]	5.7	72.2	Manual	-
TNT-Ti [Han *et al.*, 2021]	6.1	73.9	Manual	-
ViT-Ti [Dosovitskiy *et al.*, 2021]	5.7	74.5	Manual	-
CPVT-Ti [Chu *et al.*, 2021]	6	74.9	Manual	-
PVT-Tiny [Wang *et al.*, 2021]	13.2	75.1	Manual	-
ViTAS-C [Su *et al.*, 2022]	5.6	74.7	One-Shot	32
AutoFormer-Ti [Chen *et al.*, 2021b]	5.7	74.7	One-Shot	24
TF-TAS [Zhou *et al.*, 2024]	6.2	75.3	Zero-Shot	0.5
Auto-Prox [Wei *et al.*, 2024a]	6.4	75.6	Zero-Shot	0.1
Ours	9.9	**75.7**	Zero-Shot	**0.005**

Table 5: Comparison of Top-1 accuracy and search cost (GPU Days) on ImageNet within the AutoFormer search space.

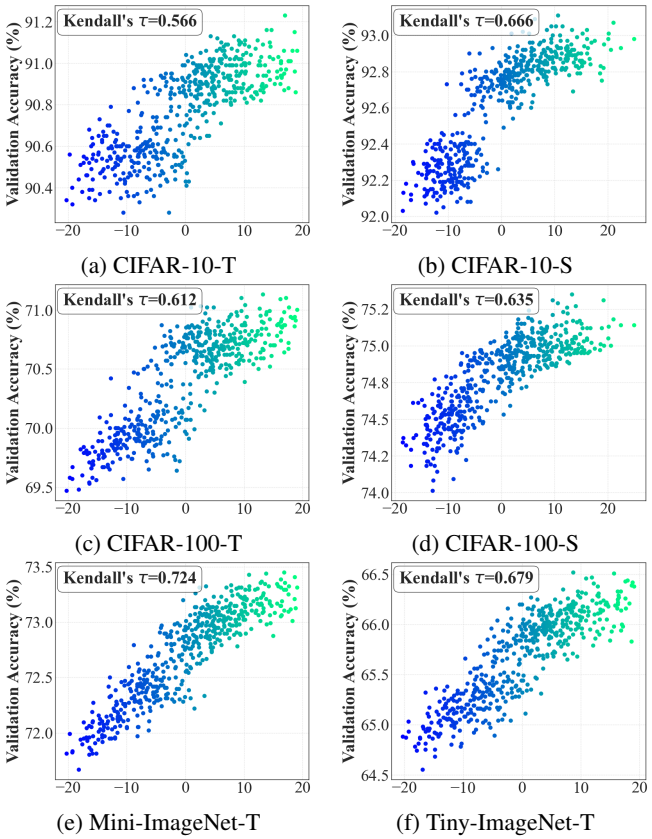
Table 2 compares the Top-1 validation accuracy and computational overhead of various NAS methods. Data-driven approaches like GraSP and SNIP suffer from high complexity, requiring over 800 ms per evaluation, rendering them impractical for large-scale search. While advanced methods such as Auto-Prox reduce latency to approximately 90 ms, they remain constrained by the necessity of forward propagation. Conversely, Info-NAS bypasses forward propagation entirely, reducing the proxy time to just 1.8 ms. This efficiency allows Info-NAS to maximize search speed while preserving high accuracy, underscoring its superiority in ViT architecture search and providing a fresh perspective on fully zero-cost proxies.

4.3 Experimental Results On ViT-Bench-101

To validate the generalization capability of Info-NAS across diverse data distributions, we compared its performance on the ViT-Bench-101 benchmark, as presented in Table 3. The results indicate that Info-NAS achieves the highest Spearman correlation coefficients on both the CIFAR-100 and Flowers datasets, maintaining remarkable ranking consistency. On the Chaoyang dataset, Info-NAS also demonstrates highly competitive performance. This consistent stability confirms that Info-NAS is a robust and efficient zero-cost proxy applicable to a wide range of ViT architecture search scenarios. Moreover, these findings further reveal a universal intrinsic connection between the internal information flow patterns of neural networks and their performance.

4.4 Experimental Results On OoD-ViT-NAS

To further validate the generalizability of Info-NAS, we evaluated it on the challenging OoD-ViT-NAS benchmark. Table 4 reports the correlation results on ImageNet-1K, while



(a) CIFAR-10-T (b) CIFAR-10-S (c) CIFAR-100-T (d) CIFAR-100-S (e) Mini-ImageNet-T (f) Tiny-ImageNet-T

Figure 4: Visualization of the correlation between Info-NAS scores and validation accuracy on various datasets.

Table 5 compares the best Info-NAS model against manually designed networks and advanced NAS methods. The results demonstrate that Info-NAS significantly outperforms existing zero-cost proxies across three correlation metrics and achieves the highest Top-1 accuracy of 75.7%. Notably, Info-NAS achieves superior search efficiency, requiring only 0.005 GPU days. This represents an order-of-magnitude speedup over comparable methods, significantly reducing computational overhead.

4.5 Ablation Study

As shown in Table 6, $\mathcal{V}$ plays a dominant role in performance evaluation, and its individual correlation significantly outperforms the other components across all datasets. This finding strongly corroborates that the total information volume of the architecture is the primary determinant of feature representation richness. In contrast, $\mathcal{G}$ and $\mathcal{C}$ function primarily as auxiliary components: $\mathcal{G}$ captures the activity of inter-layer information flow, while $\mathcal{C}$ constrains the global consistency of information transmission paths. Together, they effectively refine the capacity estimation provided by $\mathcal{V}$. Ultimately, the fusion of all components ($\mathcal{V}+\mathcal{G}+\mathcal{C}$) yields optimal performance across all benchmarks, demonstrating that our method comprehensively captures the complex architectural characteristics of ViT and enables a precise information quantification of model potential.

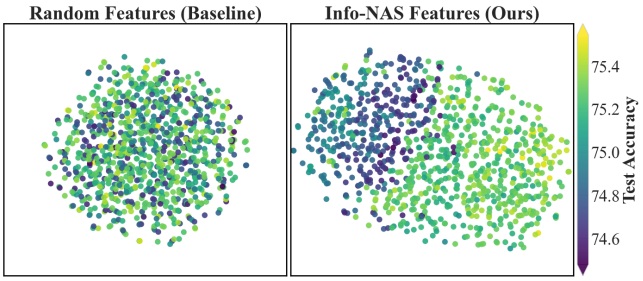


Figure 5: Feature space visualization. The color represents the test accuracy. Unlike the chaotic distribution of random features (Left), our Info-NAS features (Right) show distinct clustering consistent with model performance, validating their representational power.

Components	CIFAR-10	CIFAR-100	ImageNet
$\mathcal{G}$	0.295	0.340	0.396
$\mathcal{C}$	0.284	0.333	0.403
$\mathcal{V}$	0.540	0.620	0.774
$\mathcal{G}+\mathcal{C}$	0.307	0.360	0.434
$\mathcal{V}+\mathcal{G}$	0.538	0.617	0.766
$\mathcal{V}+\mathcal{C}$	0.558	0.632	0.774
$\mathcal{V}+\mathcal{G}+\mathcal{C}$	0.566	0.635	0.783

Table 6: Ablation Study of Different Component Combinations on CIFAR-10, CIFAR-100, and ImageNet.

4.6 Visualization of Feature Space

To validate the representational power of Info-NAS, we visualize the architecture embeddings using t-SNE [Maaten and Hinton, 2008]. Specifically, we construct feature vectors by aggregating layer-wise modulated information $\{MI_i\}_{i=1}^{L_{max}}$, padding them with zeros to the maximum network depth to ensure alignment, followed by standardization and dimensionality reduction. As shown in Figure 5, unlike the chaotic distribution of the random features, Info-NAS exhibits distinct clustering where similarly performing architectures are spatially proximal. This comparison strongly validates it as a robust feature representation for ViT performance evaluation.

5 Conclusion

In this paper, we propose Info-NAS, a zero-cost proxy approach based on information quantification, presenting the first information quantification of ViT architectures. We establish an efficient proxy evaluation framework spanning three components to facilitate the rapid assessment of ViT performance. Furthermore, we introduce the ViT-Info-Bench dataset to provide data support for the correlation analysis of zero-cost proxies. Extensive experiments across multiple benchmarks demonstrate that Info-NAS significantly outperforms existing zero-cost proxy methods in both ranking correlation and computational efficiency. It maintains high performance while substantially reducing proxy calculation overhead and search time, with each evaluation requiring only 1.8 ms. We hope this perspective, grounded in information flow, offers a promising direction for future research on efficient, training-free NAS methodologies.

Acknowledgments

This work was supported in part by the Regional Innovation and Development Joint Fund of the National Natural Science Foundation of China under Grant U25A20435, in part by the Key Research and Development Program of Hubei Province, China under Grant 2025BAB051, and in part by the Provincial Key Science and Technology Project of Hubei Province, China under Grant 2025BEA006.

Contribution Statement

[†]Yue Yang and Jiacheng Wang contributed equally.

References

[Abdelfattah *et al.*, 2021] Mohamed S. Abdelfattah, Abhinav Mehrotra, Lukasz Dudziak, and Nicholas Donald Lane. Zero-cost proxies for lightweight NAS. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Abdi, 2007] Hervé Abdi. The kendall rank correlation coefficient. *Encyclopedia of measurement and statistics*, 2:508–510, 2007.

[Alabdulmohsin *et al.*, 2023] Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36:16406–16425, 2023.

[Chen *et al.*, 2021a] Boyu Chen, Peixia Li, Chuming Li, Baopu Li, Lei Bai, Chen Lin, Ming Sun, Junjie Yan, and Wanli Ouyang. Glit: Neural architecture search for global and local image transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12–21, 2021.

[Chen *et al.*, 2021b] Minghao Chen, Houwen Peng, Jianlong Fu, and Haibin Ling. Autoformer: Searching transformers for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12270–12280, 2021.

[Chen *et al.*, 2021c] Wuyang Chen, Xinyu Gong, and Zhangyang Wang. Neural architecture search on imagenet in four gpu hours: A theoretically inspired perspective. In *International Conference on Learning Representations (ICLR)*, 2021.

[Chu *et al.*, 2021] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021.

[Dong *et al.*, 2021] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pages 2793–2803. PMLR, 2021.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Elsken *et al.*, 2019] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *J. Mach. Learn. Res.*, 20:55:1–55:21, 2019.

[Gong and Wang, 2022] Chengyue Gong and Dilin Wang. Nasvit: Neural architecture search for efficient vision transformers with gradient conflict-aware supernet training. *ICLR Proceedings 2022*, 2022.

[Guo *et al.*, 2020] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. In *European Conference on Computer Vision*, pages 544–560, 2020.

[Ha *et al.*, 2023] Hyeonjeong Ha, Minseon Kim, and Sung Ju Hwang. Generalizable lightweight proxy for robust nas against diverse perturbations. *Advances in Neural Information Processing Systems*, 36:38611–38623, 2023.

[Han *et al.*, 2021] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in neural information processing systems*, 34:15908–15919, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Ho *et al.*, 2024] Sy-Tuyen Ho, Tuan Van Vo, Somayeh Ebrahimkhani, and Ngai-Man Cheung. Vision transformer neural architecture search for out-of-distribution generalization: benchmark and insights. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 84197–84245, 2024.

[Hosseini *et al.*, 2021] Ramtin Hosseini, Xingyi Yang, and Pengtao Xie. Dsrna: Differentiable search of robust neural architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6196–6205, 2021.

[Jiang *et al.*, 2023] Tangyu Jiang, Haodi Wang, and Rongfang Bie. Meco: zero-shot nas with one data and single forward pass via minimum eigenvalue of correlation. *Advances in Neural Information Processing Systems*, 36:61020–61047, 2023.

[Lee and Ham, 2024] Junghyup Lee and Bumsub Ham. Aznas: Assembling zero-cost proxies for network architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5893–5903, 2024.

[Lee *et al.*, 2019] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Snip: single-shot network pruning

based on connection sensitivity. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2022] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[Mellor *et al.*, 2021] Joe Mellor, Jack Turner, Amos Storkey, and Elliot J Crowley. Neural architecture search without training. In *International conference on machine learning*, pages 7588–7598. PMLR, 2021.

[Noci *et al.*, 2022] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

[Peng *et al.*, 2020] Houwen Peng, Hao Du, Hongyuan Yu, Qi Li, Jing Liao, and Jianlong Fu. Cream of the crop: Distilling prioritized paths for one-shot neural architecture search. *Advances in neural information processing systems*, 33:17955–17964, 2020.

[Ren *et al.*, 2021] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys (CSUR)*, 54(4):1–34, 2021.

[Stephanou and Varughese, 2021] Michael Stephanou and Melvin Varughese. Sequential estimation of spearman rank correlation using hermite series estimators. *Journal of Multivariate Analysis*, 186:104783, 2021.

[Su *et al.*, 2022] Xiu Su, Shan You, Jiyang Xie, Mingkai Zheng, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Vitas: Vision transformer architecture search. In *European Conference on Computer Vision*, pages 139–157. Springer, 2022.

[Tanaka *et al.*, 2020] Hidenori Tanaka, Daniel Kunin, Daniel L Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in neural information processing systems*, 33:6377–6389, 2020.

[Touvron *et al.*, 2021a] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[Touvron *et al.*, 2021b] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021.

[Wang *et al.*, 2020] Chaoqi Wang, Guodong Zhang, and Roger B. Grosse. Picking winning tickets before training by preserving gradient flow. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[Wang *et al.*, 2021] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

[Wei *et al.*, 2024a] Zimian Wei, Peijie Dong, Zheng Hui, Anggeng Li, Lujun Li, Menglong Lu, Hengyue Pan, and Dongsheng Li. Auto-prox: Training-free vision transformer architecture search via automatic proxy discovery. In *AAAI*, 2024.

[Wei *et al.*, 2024b] Zimian Wei, Hengyue Pan, Lujun Li, Peijie Dong, and Dongsheng Li. Tvt: Training-free vision transformer search on tiny datasets. In *International Conference on Pattern Recognition*, pages 366–380. Springer, 2024.

[Zhai *et al.*, 2022] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

[Zhou *et al.*, 2021] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

[Zhou *et al.*, 2024] Qinqin Zhou, Kekai Sheng, Xiawu Zheng, Ke Li, Yonghong Tian, Jie Chen, and Rongrong Ji. Training-free transformer architecture search with zero-cost proxy guided evolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6525–6541, 2024.