# Auto-Prox: Training-Free Vision Transformer Architecture Search via Automatic Proxy Discovery

**Zimian Wei[1]*, Lujun Li[2]*, Peijie Dong[3], Zheng Hui[4], Anggeng Li[5], Menglong Lu[1], Hengyue Pan[1]†, Zhiliang Tian[1], Dongsheng Li[1]†**

[1]National University of Defense Technology, [2]HKUST, [3]HKUST(GZ), [4]Columbia University, [5]Huawei
{weizimian16,lumenglong,hengyuepan,tianzhiliang, dsli}@nudt.edu.cn,{lilujunai,dongpeijie98}@gmail.com
Zh2483@columbia.edu,anggeng.li@outlook.com

## Abstract

The substantial success of Vision Transformer (ViT) in computer vision tasks is largely attributed to the architecture design. This underscores the necessity of efficient architecture search for designing better ViTs automatically. As training-based architecture search methods are computationally intensive, there's a growing interest in training-free methods that use zero-cost proxies to score ViTs. However, existing training-free approaches require expert knowledge to manually design specific zero-cost proxies. Moreover, these zero-cost proxies exhibit limitations to generalize across diverse domains. In this paper, we introduce Auto-Prox, an automatic proxy discovery framework, to address the problem. First, we build the ViT-Bench-101, which involves different ViT candidates and their actual performance on multiple datasets. Utilizing ViT-Bench-101, we can evaluate zero-cost proxies based on their score-accuracy correlation. Then, we represent zero-cost proxies with computation graphs and organize the zero-cost proxy search space with ViT statistics and primitive operations. To discover generic zero-cost proxies, we propose a joint correlation metric to evolve and mutate different zero-cost proxy candidates. We introduce an elitism-preserve strategy for search efficiency to achieve a better trade-off between exploitation and exploration. Based on the discovered zero-cost proxy, we conduct a ViT architecture search in a training-free manner. Extensive experiments demonstrate that our method generalizes well to different datasets and achieves state-of-the-art results both in ranking correlation and final accuracy. Codes can be found at https://github.com/lilujunai/Auto-Prox-AAAI24.

## Introduction

Recently, Vision Transformer (ViT) (Dosovitskiy et al. 2020a) has achieved remarkable performance in image classification (Liang et al. 2022; Jiang et al. 2021; Chen, Fan, and Panda 2021), object detection (Wu et al. 2022), semantic segmentation (Dong et al. 2021), and other computer vision tasks (Li et al. 2021b; Liu et al. 2021). Despite these advancements, the manual trial-and-error method of designing ViT architectures becomes impractical given the expanding neural architecture design spaces and intricate application scenarios (Liu et al. 2023; Li et al. 2023b; Li and Jin 2022; Li et al. 2023a, 2022d,c; Li 2022; Shao et al. 2023). Neural Architecture Search (NAS) aims to address this issue by automating the design of neural network architectures. Traditional training-based architecture search methods (Xie et al. 2019; Wei et al. 2023; Hu et al. 2021; Dong et al. 2022; Chen et al. 2022; Dong, Li, and Wei 2023; Dong et al. 2023; Lu et al. 2024; Zimian Wei et al. 2024) involve training and evaluating numerous candidate ViTs, which can be computationally expensive and time-consuming. Therefore, there is a need for a more efficient architecture search of ViT.

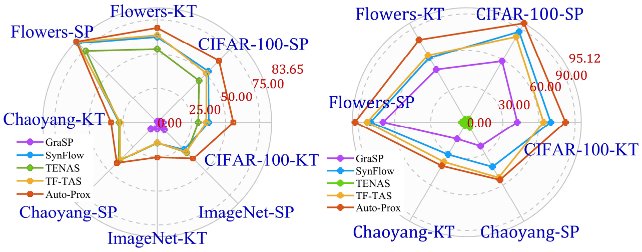

Figure 1: Kendall (KT) & Spearman (SP) ranking correlations of zero-cost proxies on AutoFormer (Left) and PiT (Right) search space for four datasets including CIFAR-100, Flowers, Chaoyang, and ImageNet. Results demonstrate that our proposed Auto-Prox significantly outperforms Synflow (Tanaka et al. 2020) and TF-TAS (Zhou et al. 2022).

Recent training-free NAS methods, such as NWOT (Mellor et al. 2021) and TF-TAS (Zhou et al. 2022), have received great research interest due to their meager costs. These methods utilize hand-crafted zero-cost proxies (Tanaka et al. 2020; Chen, Gong, and Wang 2020), which are conditional on the model's parameters or gradients, to predict the actual accuracy ranking without the expensive training process. However, there are still some drawbacks limiting their broader application: **(1) Dependency on expert knowledge and extensive tuning.** Lots of traditional zero-cost proxies are transferred from different areas with extensive expert intuition and time-consuming tuning processes. In addition, these hand-crafted zero-cost proxies can be influenced by human biases and limited by the designer's experience. **(2) Generality and flexibility.** Hand-crafted zero-cost proxies may perform well on the specific problem but can not gen-

*These authors contributed equally.

†Corresponding author