ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094709

RD-NAS: ENHANCING ONE-SHOT SUPERNET RANKING ABILITY VIA RANKING DISTILLATION FROM ZERO-COST PROXIES

Peijie Dong[1], *Xin Niu*[1,*], *Lujun Li*[2], *Zhiliang Tian*[1,*], *Xiaodong Wang*[1]
Zimian Wei[1], *Hengyue Pan*[1], *Dongsheng Li*[1]

[1] School of Computer, National University of Defense Technology, Hunan, China
[2] Chinese Academy of Sciences, Beijing, China

ABSTRACT

Neural architecture search (NAS) has made tremendous progress in the automatic design of effective neural network structures but suffers from a heavy computational burden. One-shot NAS significantly alleviates the burden through weight sharing and improves computational efficiency. Zero-shot NAS further reduces the cost by predicting the performance of the network from its initial state, which conducts no training. Both methods aim to distinguish between "good" and "bad" architectures, i.e., ranking consistency of predicted and true performance. In this paper, we propose Ranking Distillation one-shot NAS (RD-NAS) to enhance ranking consistency, which utilizes zero-cost proxies as the cheap teacher and adopts the margin ranking loss to distill the ranking knowledge. Specifically, we propose a margin subnet sampler to distill the ranking knowledge from zero-shot NAS to one-shot NAS by introducing Group distance as margin. Our evaluation of the NAS-Bench-201 and ResNet-based search space demonstrates that RD-NAS achieve 10.7% and 9.65% improvements in ranking ability, respectively. Our codes are available at https://github.com/pprp/CVPR2022-NAS-competition-Track1-3th-solution

Index Terms— neural architecture search, one-shot NAS, zero-shot NAS, rank consistency

1. INTRODUCTION

Neural Architecture Search (NAS) has sparked increased interest due to its remarkable progress in a variety of computer vision tasks [5, 3]. It aims to reduce the cost of human efforts in manually designing network architectures and discover promising models automatically. Early NAS works [6, 7] take thousands of GPU hours in search cost, and ENAS [8] first attempt at weight-sharing techniques to accelerate the searching process. The key to weight-sharing-based methods is an over-parameterized network, *a.k.a.* supernet, that encompasses all candidate architectures in the search space. The weight-sharing-based NAS approaches [1, 9] are called one-shot NAS since it only requires the cost of training one supernet. Despite the high efficiency of one-shot NAS, it is theoretically based on the ranking consistency assumption, *a.k.a.* the estimated performance of candidate architectures in supernet should be highly correlated with the true performance of the corresponding architecture when trained from scratch. However, due to the nature of the weight-sharing approach, the subnet architectures interfere with each other, and the estimated accuracy from the supernet is inevitably averaged.

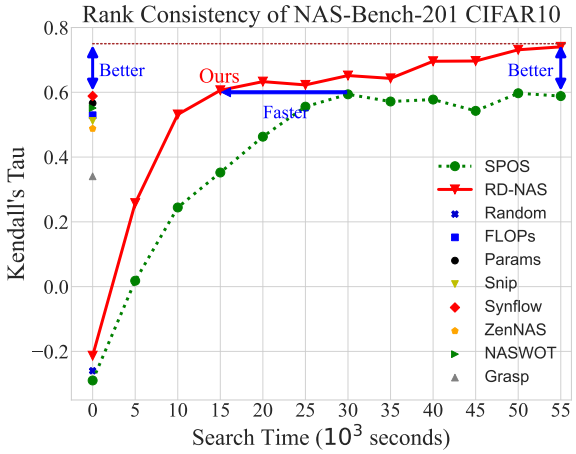
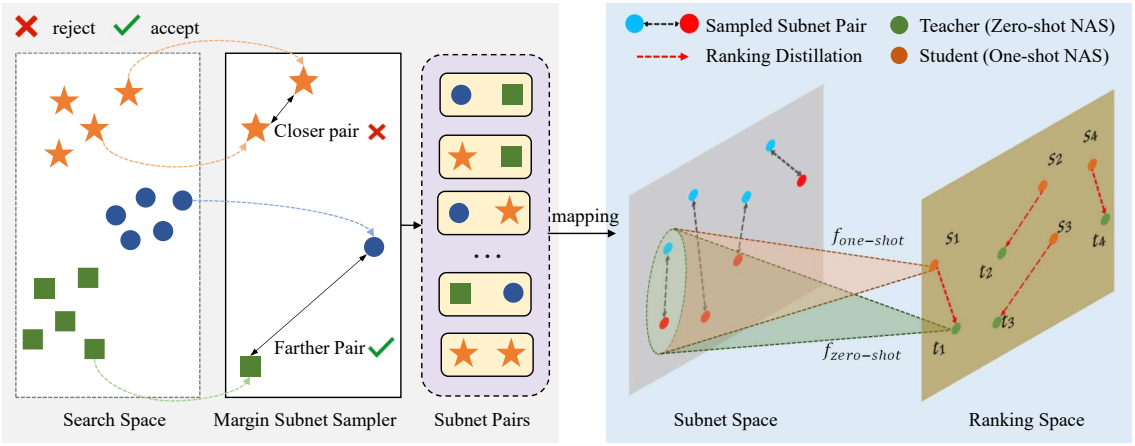


Fig. 1: The ranking consistency of one-shot NAS (SPOS [1]) and zero-shot NAS(FLOPs [2], Params [2], Snip [2], Synflow [2], ZenNAS [3], NASWOT [4], Grasp [2]). Our proposed RD-NAS achieves higher Kendall's Tau than both one-shot and zero-shot NAS and converges faster than the baseline [1].

A recent trend of NAS focuses on zero-cost proxies [4, 2], which aims to estimate the relative performance of candidate architecture from a few mini-batch of data without training the model. The approach of inferring the trained accuracy directly from the initial state of the model is called zero-shot NAS since it does not involve any training. However, there is a clear preference for zero-shot NAS that affects its ability to find good architectures [10].

To alleviate the ranking disorder caused by weight-sharing, some predictor-based methods [11] use the ranking loss to enhance the ranking ability of the model, but the challenge is that acquiring training samples (pairs of (model, accuracy)) is computationally expensive. To improve the ranking consistency

*Corresponding author



(a) Subnet Sample with Margin Subnet Sampler (b) Ranking Distillation

Fig. 2: Overview of the proposed RD-NAS. (a) Margin subnet sampler is utilized to measure the distance of subnets in the search space and sample subnet pair with margin. (b) The subnet pairs sampled in the subnet space are mapped to the ranking space by the one-shot and Zero-shot methods, and the ranking knowledge from zero-shot NAS is transferred to the one-shot NAS via ranking distillation.

economically, we resort to zero-shot NAS as an inexpensive teacher and measure the ranking relationship between a pair of subnets without employing a predictor. In recent years, the experimental investigations [2] enhanced the existing NAS algorithms using zero-cost proxies to initialize the search algorithm at the beginning of the search process, which accelerated the convergence speed of the supernet. Unlike its limited use, we incorporate zero-cost proxies into the whole training procedure with ranking knowledge distillation [12, 13].

We observe the complementarity between one-shot NAS and zero-shot NAS in Fig.1. Zero-cost proxies show good initial ranking consistency, but the upper-performance limit is not as high as one-shot NAS, while one-shot NAS has lower ranking performance in the early stage. This inspires us to bridge one-shot NAS and zero-shot NAS by utilizing zero-cost proxies as ranking teachers. In this paper, we propose a novel Ranking Distillation one-shot NAS (RD-NAS) to transfer the ranking knowledge from Zero-cost proxies to one-shot supernet.

Our contributions are summarized as follows: (1) We propose a ranking distillation framework via margin ranking loss to distill knowledge from zero-cost proxies and improve the ranking ability of one-shot NAS. (2) We propose a margin subnet sampler to reduce the uncertainty caused by zero-cost proxies by introducing group distance. (3) We demonstrate the effectiveness of RD-NAS on two types of benchmark, *a.k.a.* ResNet-like search space, and NAS-Bench-201, which achieve 10.7% and 9.65% improvement on Spearman and Kendall's Tau, respectively.

2. RANKING DISTILLATION ONE-SHOT NAS

Ranking Distillation one-shot NAS is a framework for distilling ranking knowledge from zero-shot NAS to improve the ranking ability of one-shot NAS. In Sec.2.1, we introduce how to measure the effectiveness of zero-cost proxies in the context of ranking and introduce a new criteria. In Sec.2.2, we propose the margin subnet sampler with margin ranking loss to perform the ranking distillation.

2.1. How to measure Zero-cost Proxies

Given a candidate architecture $\alpha_i \in \Omega, (i = 1, 2, ..., N)$ in the search space Ω, zero-cost proxies can quickly estimate the relative score $\pi_i^z \in \mathbb{R}$ of the i-th architecture, which can obtain the relative ranking relationship with other counterparts. However, there is a certain gap between the score π_i^z obtained by zero-cost proxies and truth scores $\pi_i^* \in \mathbb{R}$ obtained by stand-alone training. We introduce σ_i to estimate the variance of the zero-cost score of the i-th architecture, and the equation $\pi_i^* = \pi_i^z + \sigma_i$ holds. We expect to introduce truth ranking correlation between architectures into the supernet optimization process. To measure the estimation ability of zero-cost proxies, we incorporate the objective function:

$$\begin{aligned} max_{i,j} P(\pi_i^* > \pi_j^*) &= max_{i,j} P(\pi_i^z - \pi_j^z + \sigma_i - \sigma_j > 0) \\ &\approx max_{i,j} P(\pi_i^z - \pi_j^z > 0) \end{aligned} \tag{1}$$

The variance of a certain zero-cost proxy can be considered as a constant scalar, so that $\mathbb{E}[\sigma_i - \sigma_j] \approx 0$, where we use zero-cost score π^z to estimate the ground truth π^*. Here we simplify Kendall's tau and introduce Concordant Pair Ratio (CPR) δ to evaluate the pair-wise ranking consistency. We define the subnet pairs that satisfying $(\pi_i^* - \pi_j^*)(\pi_i^z - \pi_j^z) > 0$ as the concordant pair. The Concordant Pair Ratio coefficient δ is defined as:

$$\delta = \frac{\sum_{i,j} 1((\pi_i^* - \pi_j^*)(\pi_i^z - \pi_j^z) > 0)}{\sum_{i,j} 1} \tag{2}$$

which is in the range $0 \leq \delta \leq 1$. The closer δ is to 1, the better estimation of the zero-cost proxy for ground truth and *vice versa*.

Table 1: The comparison of Kendall's Tau w.r.t different methods on ResNet-like search space. "MSS" denotes the margin subnet sampler.

Type	Method	Spearman $r(\%)$
Zero-cost Proxies	Params [2]	67.23
	FLOPs [2]	78.43
	ZenNAS [3]	81.27
One-shot NAS	SPOS [1]	72.49
	Sandwich [15]	74.86
	RD-NAS w/o MSS	81.14
	RD-NAS w/ MSS	**83.20**

2.2. Ranking Distillation

To transfer the ranking knowledge obtained from zero-cost proxies to one-shot supernet, we propose a ranking distillation to further improve the ranking consistency of the one-shot NAS. In Sec.2.2.1, we introduce the margin ranking loss to distill the ranking knowledge from zero-cost proxies, but the margin is hard to set in one-shot training. Therefore, we propose a margin subnet sampler to sample subnet pairs from the subnet space in Sec.2.2.2.

2.2.1. Margin Ranking Loss

Margin ranking loss is adopted to constrain the optimization of the supernet from the perspective of ranking distillation. Unlike the standard pairwise ranking loss [14], we do not have the true rank, only the estimated rank provided by zero-cost proxies, which suffer from uncertainty. Margin ranking loss can handle difficult subnet pairs by introducing a fixed penalty margin m. The margin ranking loss of random-sampled subnet pair (α_i, α_j) is as follows:

$$\mathcal{L}^{rd}_{(\alpha_i,\alpha_j)}(\theta^s) = \max\left[0, m - \left(\mathcal{L}^{ce}\left(x, \theta^s_{\alpha_i}\right) - \mathcal{L}^{ce}\left(x, \theta^s_{\alpha_j}\right)\right)\right] \quad (3)$$

where θ^s denotes the supernet in one-shot NAS, and $\mathcal{L}^{ce}(\cdot)$ denotes the cross-entropy loss. However, such a fixed learning objective is not realistic as the training goes with inconsistent variation, which might limit the discriminative training of supernet. Instead of using a fixed margin m, we proposed a margin subnet sampler to control the margin in the subnet sampling process.

2.2.2. Margin Subnet Sampler

It is difficult to tune using a fixed margin, and the margin changes as training progresses. The effectiveness of margin ranking loss is greatly affected by the subnet pair chosen. To tackle the problems, we propose a margin subnet sampler, which affects the margin by controlling the sampled subnets. Intuitively, we expect the sampled subnet pairs to cover the entire search space, which can improve the generalization ability. We first randomly select n pairs of subnet $\{(\alpha_i, \alpha_j)\}_n$ from the search space, then we calculate the distance of the subnet pair with distance function $f(\cdot)$ mapping from subnet space $\mathbb{R}^+$ to ranking space $\mathbb{R}$ and filter out the c pairs of subnet $\{(\alpha_i, \alpha_j)\}_c = \{(\alpha_i, \alpha_j) | f(\alpha_i, \alpha_j) > \mu\}$, where distances less than the threshold μ is selected.

Table 2: Ranking correlation on NAS-Bench-201 CIFAR-10. τ, r, ρ, and δ refers to Kendall's Tau, Pearson, Spearman and CPR, respectively. "Random" refers to the rank correlation on random initialized model.

	$\tau(\%)$	$r(\%)$	$\rho(\%)$	$\delta(\%)$
FLOPs [2]	52.92	47.12	72.44	78.49
Params [2]	56.64	49.14	74.70	77.49
Snip [2]	51.16	28.20	69.12	80.49
Synflow [2]	58.83	44.90	77.64	84.99
ZenNAS [3]	48.80	55.72	67.83	44.99
NASWOT [4]	55.13	86.71	74.93	70.99
Random	-26.32	-21.43	-22.48	1.99
SPOS [1]	77.43	89.55	92.73	100.0
FairNAS [16]	70.62	82.81	88.74	100.0
RD-NAS(Ours)	**87.08**	**95.00**	**97.34**	**100.0**

Concerning the subnet distance measure $f(\cdot)$, by encoding the subnet as a sequence composed of candidate choices, the Hamming distance $f_h(\cdot)$ is computed as the number of elements that are different in the two subnet sequences. However, Hamming distance treats all candidate operations as equal and ignores the variability between them. To make the distance measure more discriminative, we propose the Margin Subnet Sampler (MSS) with Group distance $f_g(\cdot)$, which divides the candidate operations into groups with and without parameters. The intra-group distance is set small, and the inter-group distance is set large.

$$f^k_g(\alpha^k_i, \alpha^k_j) = \begin{cases} 2f_h(\alpha^k_i, \alpha^k_j), & \beta(\alpha^k_i, \alpha^k_j) = 0 \\ 0.5f_h(\alpha^k_i, \alpha^k_j), & \beta(\alpha^k_i, \alpha^k_j) = 1 \end{cases} \quad (4)$$

where $\beta(\cdot)$ is an indicator function of whether (α^k_i, α^k_j) are intra-group and $f^k_g(\cdot)$ denotes the group distance of the k-th operation of the subnet encoding.

3. EXPERIMENTS

The experiments in Sec.3.1 and Sec.3.2 verify the effectiveness of the RD-NAS on the ResNet-like search space and NAS-Bench-201. Then, we conduct the ablation study for RD-NAS.

3.1. Searching on ResNet-like Search Space

Setup. Presented in CVPR NAS Workshop, the ResNet-like search space is based on ResNet48 on the ImageNet [22], whose depth and expansion ratio are searchable. Specifically, the search space consists of four stages with different depths {5,5,8,5}, while the candidate expansion ratio is ranging from 0.7 to 1.0 at 0.05 intervals, and the channel number should be divided by 8. The search space size is 5.06×10^{19} in total. Our proposed RD-NAS won the third place at CVPR 2022 NAS Track1. We conduct a comprehensive comparison of the zero-shot [3] and one-shot methods [1, 16, 15], and the Spearman's ranking correlation is adopted to measure the

Table 3: Searching results on NAS-Bench-201 dataset [17]. Our method obtains 1.12%, 3.76%, and 4.18% absolute accuracy improvement compared with SPOS on the test sets of CIFAR-10, CIFAR-100, and ImageNet-16-120, respectively. Note that results for other weight-sharing methods are reported by NAS-Bench-201 benchmark [17], which only uses the labels of the dataset as the true performance of the searched architecture for a fair comparison.

Method	CIFAR-10 Acc. (%)		CIFAR-100 Acc. (%)		ImageNet-16-120 Acc. (%)	
	Validation	Test	Validation	Test	Validation	Test
GDAS [18]	90.00±0.21	93.51±0.13	**71.14±0.27**	70.61±0.26	41.70±1.26	41.84±0.90
DSNAS [19]	89.66±0.29	93.08±0.13	30.87±16.40	31.01±16.38	40.61±0.09	41.07±0.09
PC-DARTS [20]	89.96±0.15	**93.41±0.30**	67.12±0.39	67.48±0.89	40.83±0.08	41.31±0.22
NASWOT [4]	89.14±1.14	92.44±1.13	68.50±2.03	68.62±2.04	41.09±3.97	41.31±4.11
EPENAS [21]	89.90±0.21	92.63±0.32	69.78±2.44	70.10±1.71	41.73±3.60	41.92±4.25
Random	83.20±13.28	86.61±13.46	60.70±12.55	60.83±12.58	33.34±9.39	33.13±9.66
SPOS [1]	88.40±1.07	92.24±1.16	67.84±2.00	68.07±2.25	39.28±3.00	40.28±3.00
RD-NAS(Ours)	**90.44±0.27**	93.36±0.04	70.96±2.12	**71.83±1.33**	**43.81±0.09**	**44.46±1.58**
Optimal	91.61	94.37	73.49	73.51	46.77	47.31

ranking ability. We first pre-train the supernet for 90 epochs in ImageNet and then train the supernet with a learning rate of 0.001 with a batch size of 256 for 70 epochs. The SGD optimizer is adopted with a momentum of 0.9.

Results. The results of Spearman's ranking correlation are shown in Tab.1. RD-NAS could achieve significantly higher ranking correlations than both the zero-shot and one-shot NAS, especially using the margin subnet sampler. The Kendall's Tau of RD-NAS surpasses the baseline method [1] by 10.71% and surpasses the zero-cost proxy - "FLOPs", demonstrating that using zero-cost proxies as teachers for distillation can effectively mitigate the ranking disorder problem.

3.2. Searching on NAS-Bench-201

Setup. NAS-Bench-201 [17] provides the performance of 15,625 subnets on CIFAR-10, CIFAR-100, and ImageNet-16-120. In our experiments, the training settings are consistent with NAS-Bench-201. Following SPOS [1], we formulate the search space of NAS-Bench-201 [17] as a single-path supernet. The distance threshold μ is set to 6.7.

Evaluation Criteria. The correlation criteria used in NAS-Bench-201 are Pearson r, Kendall' Tau τ [23], Spearman ρ, and CPR δ. Pearson measures the linear relationship between the variables, while Kendall's Tau, Spearman, and CPR measure the monotonic relationship. The ranking correlations are computed using 200 randomly selected subnets.

Results. Our method obtains 2.04%, 3.12%, and 4.53% absolute accuracy improvement on the validation sets of CIFAR-10, CIFAR-100, and ImageNet-16-120 in Tab.3, compared with SPOS, which further certifies that RD-NAS enables to boost the ranking ability of one-shot NAS. Even compared with the recent NAS methods (e.g., PC-DARTS [20] and GDAS [18]), our method achieve comparable performance on the NAS-Bench-201. Especially for ImageNet-16-120, our RD-NAS achieves state-of-the-art performance that surpasses all other methods. The results on Tab.2 reveal that our RD-NAS achieve better ranking correlation than one-shot and zero-shot NAS.

Table 4: Ablation study of RD-NAS on NAS-Bench-201 CIFAR-10. "Random" denotes the baseline SPOS[1], "Margin" and "Hamming" denotes margin subnet sampler with Group and Hamming distance, respectively. "RD" denotes ranking distillation.

Random	Margin	Hamming	RD	Kendall' Tau(%)
✓	-	-	-	77.43
-	-	-	✓	79.37
-	-	✓	✓	84.18
-	✓	-	✓	**87.08**

3.3. Ablation Study

The ablation study of the distance measurement and training strategy on NAS-Bench-201 CIFAR-10 is shown in Tab.4. The one-shot training with uniform sampling strategy[1] is employed as the "Random" baseline. The results of ranking distillation exhibit the effectiveness of margin ranking loss. Moreover, using hamming or adaptive distance to sample subnets can further improve the rank consistency. The combination of margin subnet sampler and ranking distillation improves the Kendall's Tau from 77% to 87%.

4. CONCLUSION

In this paper, we present a novel ranking knowledge distillation for one-shot NAS, which we call RD-NAS, to mitigate the rank disorder problem. Specifically, our RD-NAS employs zero-cost proxies as teachers and transfers the ranking knowledge to one-shot supernet by margin ranking loss. To reduce the uncertainty of zero-cost proxies, we introduce the margin subnet sampler to increase the reliability of ranking knowledge. We conduct detailed experiments on ResNet-like search space and NAS-Bench-201 to demonstrate the effectiveness of RD-NAS. It is hoped that our work will inspire further research to exploit the strengths of zero-shot NAS.

5. REFERENCES

[1] 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 *ECCV*, 2020.

[2] Mohamed Saleh Abdelfattah, Abhinav Mehrotra, Lukasz Dudziak, and Nicholas D. Lane, "Zero-cost proxies for lightweight nas," *ArXiv*, vol. abs/2101.08134, 2021.

[3] Ming Lin, Pichao Wang, Zhenhong Sun, Hesen Chen, Xiuyu Sun, Qi Qian, Hao Li, and Rong Jin, "Zen-nas: A zero-shot nas for high-performance image recognition," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 337–346, 2021.

[4] Joseph Charles Mellor, Jack Turner, Amos J. Storkey, and Elliot J. Crowley, "Neural architecture search without training," *ArXiv*, vol. abs/2006.04647, 2021.

[5] Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Ken Chen, Wanli Ouyang, and Dong Xu, "Block proposal neural architecture search," *IEEE Transactions on Image Processing*, vol. 30, pp. 15–25, 2020.

[6] Barret Zoph and Quoc Le, "Neural architecture search with reinforcement learning," in *ICLR 2017 : International Conference on Learning Representations 2017*, 2017.

[7] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le, "Regularized evolution for image classifier architecture search," in *Proceedings of the aaai conference on artificial intelligence*, 2019, vol. 33, pp. 4780–4789.

[8] Hieu Pham, Melody Y Guan, Barret Zoph, Quoc V Le, and Jeff Dean, "Efficient neural architecture search via parameter sharing," *arXiv preprint arXiv:1802.03268*, 2018.

[9] Hanxiao Liu, Karen Simonyan, and Yiming Yang, "Darts: Differentiable architecture search," *arXiv preprint arXiv:1806.09055*, 2018.

[10] Xuefei Ning, Changcheng Tang, Wenshuo Li, Zixuan Zhou, Shuang Liang, Huazhong Yang, and Yu Wang, "Evaluating efficient performance estimators of neural architectures," in *NeurIPS*, 2021.

[11] Chi Hu, Chenglong Wang, Xiangnan Ma, Xia Meng, Yinqiao Li, Tong Xiao, Jingbo Zhu, and Changliang Li, "Ranknas: Efficient neural architecture search by pairwise ranking," in *EMNLP*, 2021.

[12] Liu Xiaolong, Li Lujun, Li Chao, and Anbang Yao, "Norm: Knowledge distillation via n-to-one representation matching," in *ICLR*, 2022.

[13] Lujun Li, Liang Shiuan-Ni, Ya Yang, and Zhe Jin, "Teacher-free distillation via regularizing intermediate representation," in *IJCNN*, 2022.

[14] Kaicheng Yu, René Ranftl, and Mathieu Salzmann, "Landmark regularization: Ranking guided super-net training in neural architecture search," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13718–13727, 2021.

[15] Jiahui Yu and Thomas Huang, "Autoslim: Towards one-shot architecture search for channel numbers," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[16] Xiangxiang Chu, Bo Zhang, Ruijun Xu, and Jixiang Li, "Fairnas: Rethinking evaluation fairness of weight sharing neural architecture search," *arXiv preprint arXiv:1907.01845*, 2019.

[17] Xuanyi Dong and Yi Yang, "Nas-bench-201: Extending the scope of reproducible neural architecture search," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[18] Xuanyi Dong and Yezhou Yang, "Searching for a robust neural architecture in four gpu hours," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1761–1770, 2019.

[19] Shoukang Hu, S. Xie, Hehui Zheng, C. Liu, Jianping Shi, Xunying Liu, and D. Lin, "Dsnas: Direct neural architecture search without parameter retraining," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12081–12089, 2020.

[20] Yuhui Xu, Lingxi Xie, Xiaopeng Zhang, Xin Chen, Guo-Jun Qi, Qi Tian, and Hongkai Xiong, "PC-DARTS: partial channel connections for memory-efficient architecture search," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2019.

[21] Vasco Lopes, Saeid Alirezazadeh, and Luís A. Alexandre, "Epe-nas: Efficient performance estimation without training for neural architecture search," in *ICANN*, 2021.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "ImageNet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[23] Pranab Kumar Sen, "Estimates of the regression coefficient based on kendall's tau," *Journal of the American Statistical Association*, vol. 63, pp. 1379–1389, 1968.