

# NATS-Bench: Benchmarking NAS Algorithms for Architecture Topology and Size

Xuanyi Dong, Lu Liu, Katarzyna Musial, Bogdan Gabrys

**Abstract**—Neural architecture search (NAS) has attracted a lot of attention and has been illustrated to bring tangible benefits in a large number of applications in the past few years. Architecture topology and architecture size have been regarded as two of the most important aspects for the performance of deep learning models and the community has spawned lots of searching algorithms for both of those aspects of the neural architectures. However, the performance gain from these searching algorithms is achieved under different search spaces and training setups. This makes the overall performance of the algorithms incomparable and the improvement from a sub-module of the searching model unclear. In this paper, we propose NATS-Bench, a unified benchmark on searching for both topology and size, for (almost) any up-to-date NAS algorithm. NATS-Bench includes the search space of 15,625 neural cell candidates for architecture topology and 32,768 for architecture size on three datasets. We analyze the validity of our benchmark in terms of various criteria and performance comparison of all candidates in the search space. We also show the versatility of NATS-Bench by benchmarking 13 recent state-of-the-art NAS algorithms on it. All logs and diagnostic information trained using the same setup for each candidate are provided. This facilitates a much larger community of researchers to focus on developing better NAS algorithms in a more comparable and computationally effective environment. All codes are publicly available at: <https://xuanyidong.com/assets/projects/NATS-Bench>.

**Index Terms**—Neural Architecture Search, Benchmark, Deep Learning

## 1 INTRODUCTION

THE deep learning community is undergoing a transition from hand-designed neural architectures [1], [2], [3] to automatically designed neural architectures [4], [5], [6], [7], [8]. In its early stages, the great success of deep learning was promoted by the introductions of novel neural architectures, such as ResNet [1], Inception [3], VGGNet [9], and Transformer [10]. However, manually designing one architecture requires human experts to frequently try and evaluate numerous different operation and connection options [4]. In contrast to architectures that are manually designed, those automatically found by neural architecture search (NAS) algorithms require much less human interaction and expert effort. These NAS-generated architectures have shown promising results in many domains, such as image recognition [4], [5], [6] and sequence modeling [5], [7], [8].

Recently, a variety of NAS algorithms have been increasingly proposed. While these NAS techniques are methodically designed and show promising improvements, many setups in their algorithms are different. (1) Different search space is utilized, e.g., range of macro skeletons of the whole architecture [11], [12], a different operation set for the micro cell within the skeleton [5], etc. (2) After a good architecture is selected, various strategies can be employed to train this architecture and report the performance, e.g., different data augmentation [13], [14], different regularization [11], different scheduler [15], and different selections of hyperparameters [16], [17]. (3) The validation set for testing the performance of the selected architecture is not

split in the same way [5], [8]. These discrepancies cause a problem when comparing the performance of various NAS algorithms, making it difficult to conclude their relative contributions.

In response to this challenge, NAS-Bench-101 [18] and NAS-HPO-Bench [19] were proposed. However, some NAS algorithms cannot be applied *directly* on NAS-Bench-101, and NAS-HPO-Bench only has 144 candidate architectures which may be insufficient to comprehensively evaluate NAS algorithms. NAS-Bench-1shot1 [20] reuses the NAS-Bench-101 dataset with some modification to analyse the one-shot NAS methods. The aforementioned works have mainly focused on the architecture topology<sup>1</sup>. However, the architecture size<sup>2</sup>, which significantly affects a model’s performance, is not considered in the existing benchmarks.

To enlarge the scope of these benchmarks and towards better reproducibility of NAS methods, we propose NATS-Bench with (1) a topology search space  $\mathcal{S}_t$  to be applicable for all NAS methods and (2) a size search space  $\mathcal{S}_s$  that supplements the lack of analysis for the architecture size. As shown in Figure 1, each architecture consists of a predefined skeleton with a stack of the searched cells. Each cell is represented as a densely-connected directed acyclic graph (DAG) as shown in the bottom section of Figure 1. The node represents the sum of the feature maps and

1. Some works [21], [22] use topology to indicate the connectivity pattern of architecture. In this manuscript, the terminology “architecture topology” or “topology” refers to the connection topology and the associated operation on each connection.

2. Some works [23] may use size to indicate the number of parameters of a neural network. In this manuscript, the terminology “architecture size” or “size” refers to the number of channels in each layer following [24].

• Xuanyi Dong, Lu Liu, Katarzyna Musial and Bogdan Gabrys are with School of Computer Science, University of Technology Sydney, NSW, Australia. (e-mail: Xuanyi.Dxy@gmail.com, Lu.Liu.Cs@icloud.com, Katarzyna.Musial-Gabrys@uts.edu.au, Bogdan.Gabrys@uts.edu.au)

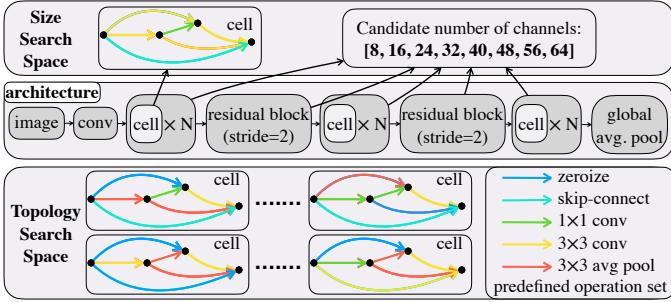


Fig. 1: **Middle:** the macro skeleton of each architecture candidate. **Top:** The size search space  $\mathcal{S}_s$  in NATS-Bench. In  $\mathcal{S}_s$ , each candidate architecture has different configuration for the channel size. **Bottom:** The topology search space  $\mathcal{S}_t$  in NATS-Bench. In  $\mathcal{S}_t$ , each candidate architecture has different cell topology.

each edge is associated with an operation transforming the feature maps from the source node to the target node.

In  $\mathcal{S}_t$ , we search for the operation assigned on each edge, and thus its size is related to the number of nodes defined for the DAG and the size of the operation set. We choose 4 nodes and 5 representative operation candidates for the operation set, which generates a total search space of 15,625 cells/architectures. In  $\mathcal{S}_s$ , we search for the number of channels in each layer (i.e., convolution, cell, or block). We pre-define 8 candidates for the number of channels, which generates a total search space of  $8^5 = 32768$  architectures. Each architecture in  $\mathcal{S}_t$  and  $\mathcal{S}_s$  is trained multiple times on three different datasets. The training log and performance of each architecture are provided for each run. The training accuracy/test accuracy/training loss/test loss after every training epoch for each architecture plus the number of parameters and floating point operations (FLOPs) are accessible.

NATS-Bench has shown its value in the field of NAS research. (1) It provides the *first* benchmark to study the architecture size. (2) It provides a unified benchmark for most up-to-date NAS algorithms including all cell-based NAS methods. With NATS-Bench, researchers can focus on designing robust searching algorithm while avoiding tedious hyperparameter tuning of the searched architecture. Thus, NATS-Bench provides a relatively fair benchmark for the comparison of different NAS algorithms. (3) It provides the full training log of each architecture. Unnecessary repetitive training procedure of each selected architecture can be avoided [4], [16] so that researchers can target on the essence of NAS, i.e., search algorithm. Another benefit is that the validation time for NAS largely decreases when testing in NATS-Bench, which provides a computational power friendly environment for more participation in NAS. (4) It provides results of each architecture on multiple datasets. The model transferability can be thoroughly evaluated for most NAS algorithms. (5) In NATS-Bench, we provide systematic analysis of the proposed search space. We also evaluate 13 recent advanced NAS algorithms including reinforcement learning (RL)-based methods, evolutionary strategy (ES)-based methods, differentiable-based methods, etc. Our empirical analysis can bring some insights to the future designs of NAS algorithms.

## 2 RELATED WORK

In the past few years, different kinds of search spaces and search algorithms have been proposed. They brought great advancements in many applications of neural network, such as visual perception [25], [26], [27], language modelling [5], [7], [8], etc. Despite their success, many researchers have raised concerns about the reproducibility and generalization ability of the NAS algorithms [?], [18], [20], [28], [29], [29]. It is essentially not clear if the reported improvements have come from hyperparameter settings, re-training pipelines, random seeds, or the improvements of the searching algorithm itself [?]. Many researchers devote their effort to solve this problem, and we will introduce them in Section 2.1 and Section 2.2.

### 2.1 NAS Benchmark

To the best of our knowledge, NAS-Bench-101 [18] is the only existing large-scale architecture dataset. Similar to NATS-Bench, NAS-Bench-101 also transforms the problem of architecture search into the problem of searching neural cells, represented as a DAG. Differently, NAS-Bench-101 defines operation candidates on the node, whereas we associate operations on the edge as inspired by [7], [8], [11]. We summarize characteristics of our NATS-Bench and NAS-Bench-101 in Table 1. The main highlights of our NATS-Bench are as follows. (1) The search spaces in NATS-Bench includes both architecture topology and size, but NAS-Bench-101 only has a topology search space. (2) NATS-Bench is algorithm-agnostic while NAS-Bench-101 without any modification is only applicable to selected algorithms [20], [30]. The original complete search space, based on the nodes in NAS-Bench-101, is huge. So, it is exceedingly difficult to efficiently traverse the training of all architectures. To trade off the computational cost and the size of the search space, they constrain the maximum number of edges in the DAG. However, it is difficult to incorporate this constraint in all NAS algorithms, such as NAS algorithms based on parameter sharing [5], [8]. Therefore, many NAS algorithms cannot be directly evaluated on NAS-Bench-101. Our NATS-Bench solves this problem by sacrificing the number of nodes and including all possible edges so that our search space is algorithm-agnostic. (3) We provide architecture information on three (instead of one) datasets and extra diagnostic information, such as architecture computational cost, fine-grained training and evaluation time, etc., which we hope will give inspirations to better and more efficient designs of NAS algorithms.

Despite the existence of NAS-Bench-101, other researchers have also devoted their effort to building a fair comparison and development environments for NAS. Zela et al. [20] proposed a general framework for one-shot NAS methods and reused NAS-Bench-101 to benchmark different NAS algorithms. Yu et al. [30] designed a novel evaluation framework to evaluate the search phase of NAS algorithms by comparing with a random search. The aforementioned works have mainly focused on the network topology. However, as other aspects of DNNs, such as network size and optimizer, significantly affect the network's performance, there is a need for an environment and systematic studies covering these areas of NAS. Unfortunately,

	#Unique Architectures	#Datasets	Diagnostic Information	Search Space	Supported NAS Algorithms			
					RL	ES	Diff.	HPO
NAS-Bench-101	423k	1	x	topology	partial	partial	none	most
$\mathcal{S}_t$ in NATS-Bench	6.5k	3	fine-grained accuracy	topology	all	all	all	most
$\mathcal{S}_s$ in NATS-Bench	32.8k	3	and loss, parameters, etc	size	all	all	most	most

TABLE 1: We summarize the important characteristics of NAS-Bench-101 and NATS-Bench. Our NATS-Bench provides the search space for both architecture topology and architecture size. Besides, NATS-Bench provides train/validation/test performance on three (one for NAS-Bench-101) different datasets so that the generality of NAS algorithms can be evaluated. It also provides some diagnostic information that may provide insights to design better NAS algorithms.

until now these aspects have rarely been considered with regard to the problem of reproducibility and generalization ability.

## 2.2 HyperParameter Optimization (HPO) Benchmark

NAS-HPO-Bench [19] evaluated 62208 configurations in the joint NAS and hyperparameter space for a simple 2-layer feed-forward network. Since NAS-HPO-Bench has only 144 architectures, it may be insufficient to evaluate different NAS algorithms. The NAS-HPO-Bench dataset also includes the number of channels in a multi-layer perceptron (MLP). In contrast, our NATS-Bench has a much larger size search space than NAS-HPO-Bench and provides the useful information on deep architecture instead of shallow MLP.

## 3 NATS-BENCH

Our NATS-Bench is algorithm-agnostic. Put simply, it is applicable to almost any up-to-date NAS algorithm. In this section, we will briefly introduce our NATS-Bench. The search space of NATS-Bench is inspired by cell-based NAS algorithms (Section 3.1). NATS-Bench evaluates each architecture on three different datasets (Section 3.2). All implementation details of NATS-Bench are introduced in Section 3.3. NATS-Bench also provides some diagnostic information which can be used for potentially better designs of future NAS algorithms (discussed in Section 3.4).

### 3.1 Architectures in the Search Space

**Macro Skeleton.** Our search space follows the design of its counterpart as used in the recent neural cell-based NAS algorithms [5], [8], [11]. As shown in the middle part of Figure 1, the skeleton is initiated with one 3-by-3 convolution with 16 output channels and a batch normalization layer [31]. The main body of the skeleton includes three stacks of cells, connected by a residual block. All cells in an architecture has the same topology. The intermediate residual block is the basic residual block with a stride of 2 [1], which serves to down-sample the spatial size and double the channels of an input feature map. The shortcut path in this residual block consists of a 2-by-2 average pooling layer with stride of 2 and a 1-by-1 convolution. The skeleton ends up with a global average pooling layer to flatten the feature map into a feature vector. The classification uses a fully connected layer with a softmax layer to transform the feature vector into the final prediction.

**The Topology Search Space  $\mathcal{S}_t$ .** The topology search space is inspired by the popular cell-based NAS

algorithms [7], [8], [11]. Since all cells in an architecture have the same topology, an architecture candidate in  $\mathcal{S}_t$  corresponds to a different cell, which is represented as a densely connected DAG. The densely connected DAG is obtained by assigning a direction from the  $i$ -th node to the  $j$ -th node ( $i < j$ ) for each edge in an undirected complete graph. Each edge in this DAG is associated with an operation transforming the feature map from the source node to the target node. All possible operations are selected from a predefined operation set, as shown in Figure 1(bottom-right). In our NATS-Bench, the predefined operation set  $\mathcal{O}$  has  $L = 5$  representative operations: (1) zeroize, (2) skip connection, (3) 1-by-1 convolution, (4) 3-by-3 convolution, and (5) 3-by-3 average pooling layer. The convolution in this operation set is an abbreviation of an operation sequence of ReLU, convolution, and batch normalization. The DAG has  $V = 4$  nodes, where each node represents the sum of all feature maps transformed through the associated operations of the edges pointing to this node. We choose  $V = 4$  to allow the search space to contain basic residual block-like cells, which require 4 nodes. Densely connected DAG does not restrict the searched topology of the cell to be densely connected, since we include zeroize in the operation set, which is an operation of dropping the associated edge. We do not impose the constraint on the maximum number of edges [18], and thus  $\mathcal{S}_t$  is applicable to most NAS algorithms, including all cell-based NAS algorithms. For each architecture in  $\mathcal{S}_t$ , each cell is stacked  $N = 5$  times, with the number of output channels set to 16, 32 and 64 for the first, second and third stages, respectively.

**The Size Search Space  $\mathcal{S}_s$ .** The size search space is inspired by transformable architecture search methods [24], [32], [33]. In the size search space, every stack in each architecture is constructed by stacking  $N = 1$  cell. All cells in every architecture have the same topology, which is the best one in  $\mathcal{S}_t$  on the CIFAR-100 dataset. Each architecture candidate in  $\mathcal{S}_s$  has a different configuration regarding the number of channels in each layer.<sup>3</sup> We build the size search space  $\mathcal{S}_s$  to include the largest number of channels in  $\mathcal{S}_t$ . Therefore, the number of channels in each layer is chosen from  $\{8, 16, 24, 32, 40, 48, 56, 64\}$ . Therefore, the size search space  $\mathcal{S}_s$  has  $8^5 = 32768$  architecture candidates.

### 3.2 Datasets

We train and evaluate each architecture on CIFAR-10, CIFAR-100 [34], and ImageNet-16-120 [35]. We choose these

<sup>3</sup> A layer could be the stem 3-by-3 convolutional layer, the cell, or the residual block.

optimizer	Nesterov	learning rate (LR)	momentum	weight decay	batch size	norm	random flip	random crop	epoch		
value	SGD	✓	cosine decay	LR from 0.1 to 0	0.9	0.0005	256	✓	p=0.5	✓	12

TABLE 2: The training hyperparameters  $\mathcal{H}^0$  for all candidate architectures in  $\mathcal{S}_s$  and  $\mathcal{S}_t$ .

three datasets because CIFAR and ImageNet [36] are the most popular image classification datasets.

We split each dataset into training, validation and test sets to provide a consistent training and evaluation settings for previous NAS algorithms [8]. Most NAS methods use the validation set to evaluate architectures after the architecture is optimized on the training set. The validation performance of the architectures serves as the supervision signals to update the searching algorithm. The test set is to evaluate the performance of each searching algorithm by comparing the indicators (e.g., accuracy, #parameters, speed) of their selected architectures. Previous methods use different splitting strategies, which may result in various searching costs and unfair comparisons. We hope to use the proposed splits to unify the training, validation and test sets for a fairer comparison.

**CIFAR-10:** It is a standard image classification dataset and consists of 60K  $32 \times 32$  colour images in 10 classes. The original training set contains 50K images, with 5K images per class. The original test set contains 10K images, with 1K images per class. Due to the need of validation set, we split all 50K training images in CIFAR-10 into two groups. Each group contains 25K images with 10 classes. We regard the first group as the new training set and the second group as the validation set.

**CIFAR-100:** This dataset is just like CIFAR-10. It has the same images as CIFAR-10 but categorizes each image into 100 fine-grained classes. The original training set on CIFAR-100 has 50K images, and the original test set has 10K images. We randomly split the original test set into two groups of equal size — 5K images per group. One group is regarded as the validation set, and another one is regarded as the new test set.

**ImageNet-16-120:** We build ImageNet-16-120 from the down-sampled variant of ImageNet (ImageNet16×16). As indicated in [35], down-sampling images in ImageNet can largely reduce the computation costs for optimal hyperparameters of some classical models while maintaining similar searching results. [35] down-sampled the original ImageNet to  $16 \times 16$  pixels to form ImageNet16×16, from which we select all images with label  $\in [1, 120]$  to construct ImageNet-16-120. In sum, ImageNet-16-120 contains 151.7K training images, 3K validation images, and 3K test images with 120 classes.

By default, in this paper, “the training set”, “the validation set”, “the test set” indicate the new training, validation, and test sets, respectively.

### 3.3 Architecture Performance

**Training Architectures.** In order to unify the performance of every architecture, we provide the performance of every architecture in our search space. In our NATS-Bench, we follow previous literature to set up the hyperparameters and training strategies [1], [11], [15]. We train each architecture

with the same strategy, which is shown in Table 2. For simplification, we denote all hyperparameters for training a model as a set  $\mathcal{H}$ . We use  $\mathcal{H}^0$ ,  $\mathcal{H}^1$ , and  $\mathcal{H}^2$  to denote the three kinds of hyperparameters that we use. Specifically, we train each architecture via Nesterov momentum SGD, using the cross-entropy loss. We set the weight decay to 0.0005 and decay the learning rate from 0.1 to 0 with a cosine annealing [15]. We use the same  $\mathcal{H}^0$  on different datasets, except for the data augmentation which is slightly different due to the image resolution. On the CIFAR datasets, we use the random flip with probability of 0.5, the random crop  $32 \times 32$  patch with 4 pixels padding on each border, and the normalization over RGB channels. On ImageNet-16-120, we use a similar strategy but with random crop  $16 \times 16$  patch and 2 pixels padding on each border. In  $\mathcal{H}^0$ , we train each architecture by 12 epochs, which can be used in bandit-based algorithms [37], [38]. Since 12 epochs are not sufficient to evaluate the relative ranking of different architectures, we train each candidate with more epochs ( $\mathcal{H}^1$  and  $\mathcal{H}^2$ ) to obtain a more accurate ranking.  $\mathcal{H}^1$  and  $\mathcal{H}^2$  are the same as  $\mathcal{H}^0$  but use 200 epochs and 90 epochs, respectively. In NATS-Bench, we apply  $\mathcal{H}^0$  and  $\mathcal{H}^1$  on the topology search space  $\mathcal{S}_t$ ; and we apply  $\mathcal{H}^0$  and  $\mathcal{H}^2$  on the size search space  $\mathcal{S}_s$ .

**Metrics.** We train each architecture with different random seeds on different datasets. We evaluate each architecture  $\alpha$  after every training epoch. NATS-Bench provides the training, validation, and test loss as well as accuracy. Users can easily use our API to query the results of each trial of  $\alpha$ , which has negligible computational costs. In this way, researchers could significantly speed up their searching algorithm on these datasets and focus solely on the essence of NAS.

### 3.4 Diagnostic Information

Validation accuracy is a commonly used supervision signal for NAS. However, considering the expensive computational costs for evaluating the architecture, the signal is too sparse. In our NATS-Bench, we also provide some additional diagnostic information in a form of extra statistics obtained during training of each architecture. Collecting these statistics almost involves no extra computation cost but may provide insights for better designs and training strategies of different NAS algorithms, such as platform-aware NAS [12], accuracy prediction [39], mutation-based NAS [40], [41], etc.

**Architecture computational costs:** NATS-Bench provides three computation metrics for each architecture — the number of parameters, FLOPs, and latency. Algorithms that focus on searching architectures with computational constraints, such as models on edge devices, can use these metrics directly in their algorithm designs without extra calculations. We also provide the training time and evaluation time for each architecture.

**Fine-grained training and evaluation information.** NATS-Bench tracks the changes in loss and accuracy of

every architecture after every training epoch. These fine-grained training and evaluation information often shows the trends related to the architecture performance and could help with identifying some attributes of the model, such as the speed of convergence, the stability, the over-fitting or under-fitting levels, etc. These attributes may benefit the designs of NAS algorithms. Besides, some methods learn to predict the final accuracy of an architecture based on the results of a few early training epochs [39]. These algorithms can be trained faster, and the performance of the accuracy prediction can be evaluated using the fine-grained evaluation information.

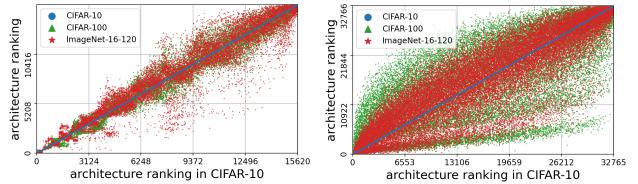
**Parameters of the optimized architecture.** Our NATS-Bench releases the trained parameters for each architecture. This can provide ground truth label for hypernetwork-based NAS methods [42], [43], which learn to generate parameters of architecture. Other methods mutate an architecture to become another one [6], [40]. With NATS-Bench, researchers could directly use the off-the-shelf parameters instead of training them from scratch and analyze how to transfer parameters from one architecture to another.

### 3.5 What/Who can Benefit from NATS-Bench?

Our NATS-Bench provides a unified NAS library for the community and can benefit NAS algorithms from the perspective of both performance and efficiency. NAS has been dominated by multi-fidelity based methods [6], [29], [37], [39], which learn to search based on an approximation of the performance of each candidate to accelerate searching. Running algorithms on our NATS-Bench can reduce the approximation to an accurate performance via only querying from the database. This can avoid sub-optimal training because of the inaccurate estimation of the performance as well as accelerate the training into seconds. Meanwhile, with the provision of our diagnostic information, such as latency, algorithms trained with such extra pieces of information can directly fetch them from our codebase with negligible efforts. Meanwhile, the designs of NAS algorithms can also have more diversity with the benefit of the diagnostic information, and more potential designs will be discussed in Section 6.

In the NAS community, there has been growing attention to the field of both searching for topology [5], [8] and searching for size [44], [45]. By benchmarking either topology or size on NATS-Bench, it may help researchers to understand the effectiveness of these algorithms and give inspirations for ongoing and future research which lies in this intersection.

NATS-Bench provides a unified codebase – a NAS library – to make the benchmarking as fair as possible. In this codebase, we share the code implementation for different algorithms as much as possible. For example, the super network for weight-sharing methods is reused; the data pipelines for different methods are reused; the interface of training, forwarding, optimizing for different algorithms is kept the same. We demonstrate, using 13 state-of-the-art NAS algorithms applied on NATS-Bench, how the process has been unified through an easy-to-use API. The implementation difference between DARTS [8] and GDAS [7] is only less than 20 lines of code. Our library reduces



(a) The relative ranking for (b) The relative ranking for  
the topology search space  $\mathcal{S}_t$ . the size search space  $\mathcal{S}_s$ .

Fig. 2: The ranking of each architecture on three datasets, sorted by the ranking in CIFAR-10.

the effect caused by the implementation difference when comparing different methods. It is also easy to implement new NAS algorithms by reusing and extending our library. More detailed engineering designs can be found in the documentation of our released codes. As this part is beyond the scope of this manuscript, we do not introduce it here.

## 4 ANALYSIS OF NATS-BENCH

### 4.1 An Overview of Architecture Performance

The performance of each architecture in both search spaces  $\mathcal{S}_t$  and  $\mathcal{S}_s$  is shown in Figure 3. The training and test accuracy with respect to the number of parameters and number of FLOPs are shown in each column, respectively. Results show that a different number of parameters or FLOPs will affect the performance of the architectures, which indicates that the choices of operations are essential in NAS. We also observe that the performance of the architecture can vary even when the number of parameters or FLOPs stays the same.

These observations indicate the importance of how the operations are connected and how the number of channels is set. We compare all architectures in  $\mathcal{S}_t$  and  $\mathcal{S}_s$  with some classical human-designed architectures (orange star marks in Figure 3). (I) Compared to candidates in  $\mathcal{S}_t$ , ResNet shows competitive performance in three datasets, however, it still has room to improve, i.e., about 2% compared to the best architecture in CIFAR-100 and ImageNet-16-120, about 1% compared to the best one with the same amount of parameters in CIFAR-100 and ImageNet-16-120. (II) In many vision tasks, the pyramid structure, where the number of channels is gradually increased [46], has shown superior generalization ability. Inspired by this, we plot three candidates that have a pyramid structure: the number of channels in each layer is 8-16-24-32-40, 8-16-32-48-64, and 32-40-48-56-64. Regarding the parameters vs. the accuracy, these pyramid candidates in  $\mathcal{S}_s$  are far from the Pareto optimality. Regarding the FLOPs vs. the accuracy, they are close to Pareto optimality.

### 4.2 Architecture Ranking on Three Datasets

The ranking of every architecture in our search space is shown in Figure 2, where the architectures ranked in CIFAR-10 (x-axis) are shown in relation to their respective ranks in CIFAR-100 and ImageNet-16-120 (y-axis), indicated by green and red markers respectively. The performance of the architectures in  $\mathcal{S}_t$  shows a generally consistent ranking over

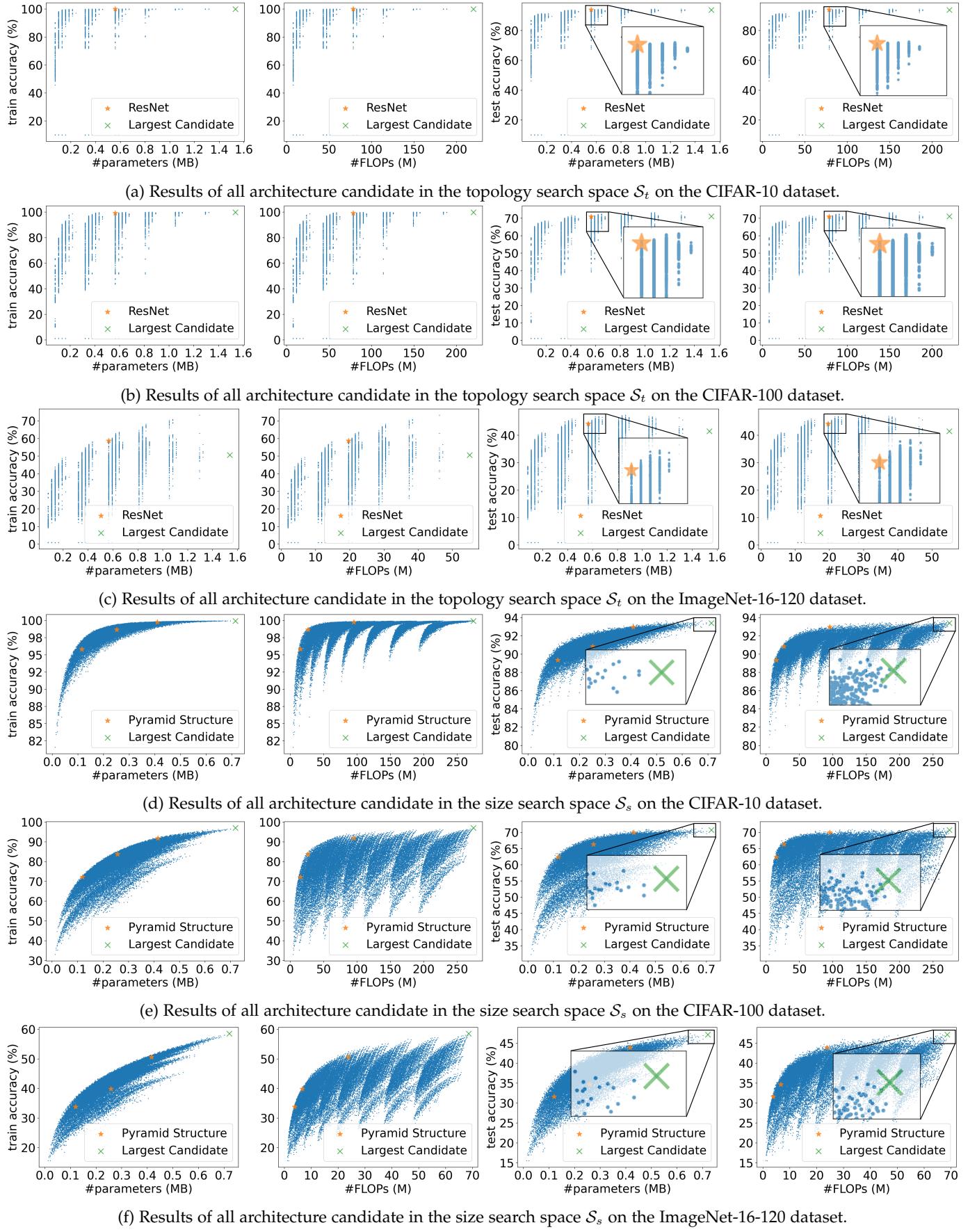
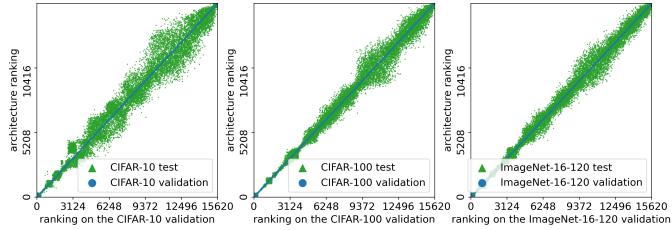
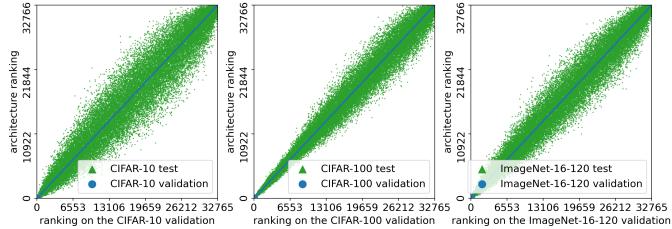


Fig. 3: The training and test accuracy vs. the number of parameters and FLOPs for each architecture candidate.

(a) The relative ranking for the topology search space  $\mathcal{S}_t$ .Fig. 4: The correlation between the validation accuracy and the test accuracy for all architecture candidates in  $\mathcal{S}_t$  and  $\mathcal{S}_s$ .

the three datasets with slightly different variance, which serves to test the generality of the searching algorithm. In contrast, the ranking of architecture candidates in  $\mathcal{S}_s$  is quite different. It indicates that the optimal architecture sizes on three datasets are different.

We compute the validation as well as the test accuracy after training with  $\mathcal{H}^1$  and  $\mathcal{H}^2$  on  $\mathcal{S}_t$  and  $\mathcal{S}_s$ , respectively. Figure 4 visualizes their correlation. It shows the relative ranking obtained from the validation accuracy is similar to that obtained using the test accuracy. Thus, it guarantees the upper bounds of the NAS algorithms as the brute-force strategy can find an architecture that can almost achieve the highest test accuracy.

We also show the Kendall rank correlation coefficient [47] across different datasets in Figure 6. This rank correlation dramatically decreases as we only pick the top performing architecture candidates. When we directly transfer the best architecture in one dataset to another (i.e. a vanilla strategy), it can not 100% secure a good performance. This phenomena is a call for better transferable NAS algorithms instead of using the vanilla strategy.

**Ranking stability of top architectures.** The accuracy of an architecture trained in different trials may have a high variance. Such variance of accuracy may affect the architecture rankings. To investigate such effect, we compare two kinds of rankings: (1) average the accuracy of multiple trials, and use the averaged accuracy to compute the architecture rankings; (2) randomly select a trial for each architecture, and use the accuracy of the randomly selected trial to compute the architecture rankings. We show these two rankings of three datasets in Figure 5. The Kendall rank correlation coefficient on CIFAR-10, CIFAR-100, and ImageNet-16-120 are about 0.77, 0.70, and 0.77, respectively.

## 5 BENCHMARK

### 5.1 Background

NAS aims to find architecture  $\alpha$  among the search space  $\mathcal{S}$  so that this found  $\alpha$  achieves a high performance on the

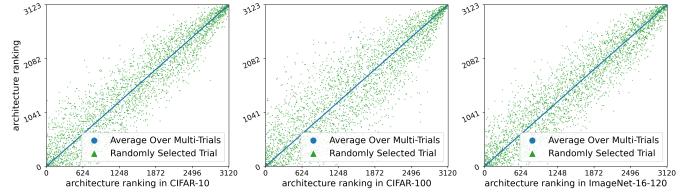
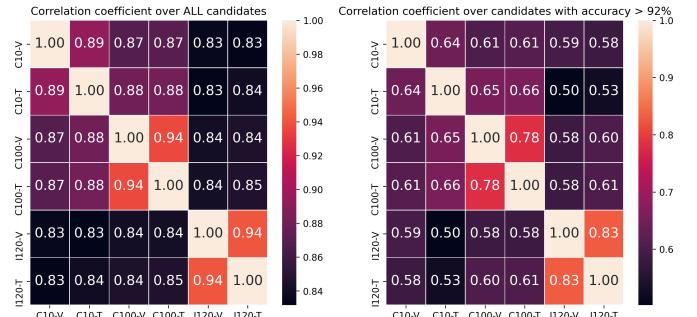
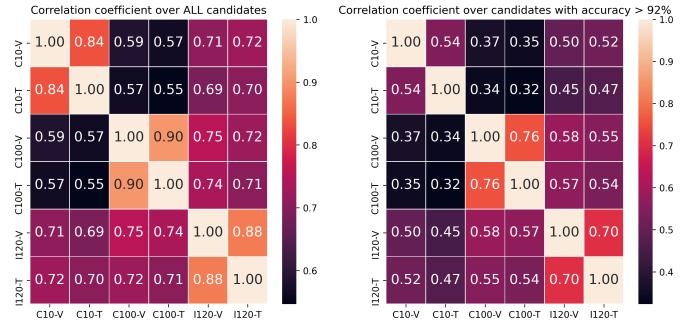
Fig. 5: Ranking stability of top 20% architectures on different datasets over the topology search space  $\mathcal{S}_t$ .(a) The Kendall rank correlation coefficient for  $\mathcal{S}_t$ .(b) The Kendall rank correlation coefficient for  $\mathcal{S}_s$ .

Fig. 6: We report the Kendall rank correlation coefficient between the accuracy on 6 sets, i.e., CIFAR-10 validation set (C10-V), CIFAR-10 test set (C10-T), CIFAR-100 validation set (C100-V), CIFAR-100 test set (C100-T), ImageNet-16-120 validation set (I120-V), ImageNet-16-120 test set (I120-T).

validation set. This problem can be formulated as a bi-level optimization problem:

$$\begin{aligned} \min_{\alpha \in \mathcal{S}} \mathcal{L}(\alpha, \omega^*, \mathcal{D}_{val}) \\ \text{s.t. } \omega^* = \arg \min_{\omega} \mathcal{L}(\alpha, \omega, \mathcal{D}_{train}), \end{aligned} \quad (1)$$

where  $\mathcal{L}$  indicates the objective function (e.g., cross-entropy loss).  $\mathcal{D}_{train}$  and  $\mathcal{D}_{val}$  denote the training data and the validation data, respectively. In the typical NAS setting, after an architecture  $\alpha$  is found,  $\alpha$  will be re-trained on  $\mathcal{D}_{train}$  (or  $\mathcal{D}_{train} + \mathcal{D}_{val}$ ) and evaluated on the test data  $\mathcal{D}_{test}$  to figure out its real performance.

### 5.2 Experimental Setup

We evaluate 13 recent, state-of-the-art searching methods on our NATS-Bench, which can serve as baselines for future NAS algorithms in our dataset. Specifically, we evaluate some typical NAS algorithms: (I) Random Search algorithms, e.g., random search (RANDOM) [48], random

Accelerate the search	RANDOM, REINFORCE, REA, and BOHB
Accelerate the evaluation	all NAS methods

TABLE 3: The utility of our NATS-Bench for different NAS algorithms. We show whether a NAS algorithm can use our NATS-Bench to accelerate the searching and evaluation procedure.

search with parameter sharing (RSPS) [29]. (II) ES methods, e.g., REA [6]. (III) RL algorithms, e.g., REINFORCE [49], ENAS [5]. (IV) Differentiable algorithms. e.g., first order DARTS (DARTS 1st) [8], second order DARTS (DARTS 2nd), GDAS [7], SETN [17], TAS [24], FBNet-V2 [45], TuNAS [50]. (V) HPO methods, e.g., BOHB [37].

Among them, RANDOM, REA, REINFORCE, and BOHB are multi-trial based methods. They can be used to search on both  $\mathcal{S}_t$  and  $\mathcal{S}_s$  search spaces. Especially, using our API, we can accelerate them to be executed in seconds as shown in Table 3.

Other methods are weight-sharing based methods, in which the evaluation procedure can be accelerated by using our API. Notably, DARTS, GDAS, SETN are specifically designed for the topology search space  $\mathcal{S}_t$ . The search strategies for #channels in TAS, FBNet-V2, and TuNAS can be used on the size search space  $\mathcal{S}_s$ .

### 5.3 Experimental Results

#### 5.3.1 Multi-trial based Methods

We follow the suggested hyperparameters in their original papers to run each method on our topology search space  $\mathcal{S}_t$  and size search space  $\mathcal{S}_s$ . We run each experiment 500 times on three datasets. For the CIFAR-10 dataset, we set up a maximum time budget of 2e4 seconds. As the training time for a single model is much larger on other datasets than that on CIFAR-10, we increase this time budget for other datasets accordingly. Every 100 seconds, each method can let us know the current searched architecture candidate. We use the hyperparameters  $\mathcal{H}^0$  (12 epochs) to obtain a validation accuracy for each trial. This validation accuracy serves as the supervision/feedback signal for these multi-trial based methods. For BOHB, given its current budget for a trial, it can early stop before fully training the model using 12 epochs. We show the averaged accuracy of this searched architecture candidate over 500 runs in Figure 7 and Table 4. Each sub-figure in Figure 7 corresponds to one dataset and a search space. For example, in the middle of Figure 7b, we search on CIFAR-100 and show the test accuracy of the automatically discovered architecture on CIFAR-100.

**Observations on the topology search space  $\mathcal{S}_t$ .** (1) On CIFAR-10, most methods have similar performance. (2) On CIFAR-100, before 2e4 seconds, REA is similar to BOHB, and they outperforms REINFORCE and RANDOM; at 4e4 seconds,  $REA \geq BOHB \geq REINFORCE \geq RANDOM$ . (3) On ImageNet-16-120, BOHB converges faster than the other methods. It may be caused by the dynamic budget mechanism for each trial in BOHB, which allows to traverse more architecture candidates.

**Observations on the size search space  $\mathcal{S}_s$ .** (1) REA significantly outperforms the other methods on all datasets in the size search space  $\mathcal{S}_s$ . (2) On CIFAR-10, REINFORCE

is better than BOHB and RANDOM. (3) On CIFAR-100 and ImageNet-16-120, the results of BOHB and REINFORCE are similar, and RANDOM is the worst one. (4) As the searching time goes, the searched architecture by REA becomes closer and closer to the best one, while the other methods need much more time to catch up with REA. (5) Figure 3 implies a simple prior for  $\mathcal{S}_s$ : without the constraint of model cost, the larger model tends to have higher accuracy. By visualising the searched architecture, REA can quickly fit this prior while the other methods do not.

Given the flexibility and robustness of REA, we would recommend choosing REA as a searching algorithm if the computational resources are sufficient.

#### 5.3.2 Weight-sharing based Methods

To compare weight-sharing based methods as fairly as possible, we keep the same hyperparameters concerned with the optimising of the shared weights for different methods. For other hyperparameters, e.g., hyperparameters for optimising the controller in ENAS or hyperparameters for optimising the architectural parameters in DARTS/GDAS, we use the same values as introduced in their original papers by default. For the NAS algorithms for the size search space, we follow [50] to warmup the one-shot model at the first 30% search phase. In this way, we can focus on evaluating the core and unique modules in each searching algorithm. We setup the total number of epochs to 100 for search, and compare results of their searched architecture candidates after each search epoch. We run each experiment three times and report the average results in Figure 8a, Figure 8b, and Table 4.

**Observations on the topology search space  $\mathcal{S}_t$ .** (1) On CIFAR-10, DARTS (1st) and DARTS (2nd) quickly converge to find the architecture having many skip connections, which performs poorly. However, on CIFAR-100 and ImageNet-16-120, they perform relatively well. This is because the significantly increased searching data on CIFAR-100 and ImageNet-16-120 over CIFAR-10 alleviate the problem of incorrect gradient estimation in bi-level optimization. (2) RSPS, ENAS, and SETN converge quickly and are robust on three datasets. During their searching procedure, they will randomly sample some architecture candidates, evaluate them using the shared weights, and select the candidate with the highest validation accuracy. Such strategy is more robust than using the arg max over the learned architecture parameters in [7], [8]. (3) The searched architecture of GDAS slowly converges to the similar one as ENAS and SETN.

Some observations on  $\mathcal{S}_t$  are different from those in our preliminary version. It is because some hyperparameters changed following either suggestions from the authors or better strategies found in our experiments. Especially, we would like to highlight some useful strategies for weight-sharing based methods: (1) always use batch statistics for the batch normalization layer. (2) use the same configuration of the standalone for the one-shot model, such as number of layers and batch size. (3) during the evaluation procedure of RSPS, ENAS, and SETN, the average accuracy for a large batch of validation data is sufficient to approximate the average accuracy on the whole validation set. In our experiments, we use the batch size of 512 for evaluation.

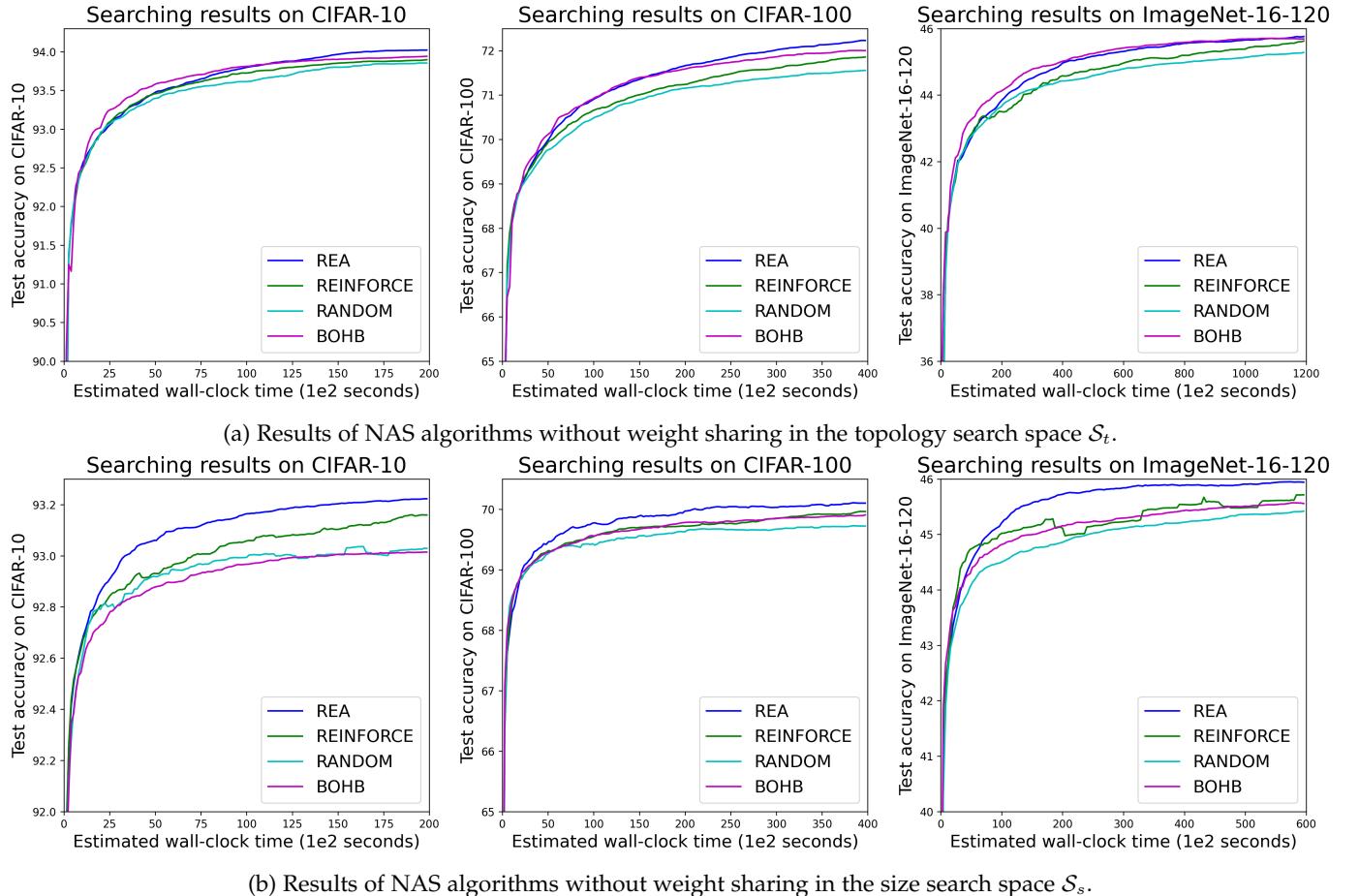


Fig. 7: The test accuracy of the searched architecture candidate over time. We run different searching algorithms 500 times on three datasets. We plot the test accuracy of their searched model at each timestamp for the corresponding dataset. This test accuracy is evaluated after fully training the model on the corresponding dataset and averaged over 500 runs. We report more details including the accurate numbers and variance in Table 4.

**Observations on the size search space  $S_s$ .** We abstract the three kinds of strategies to search for #channels:

- Using channel-wise interpolation to explicitly compare two different #channels [24].
- Using the masking mechanism to represent different candidate #channels and optimize its distribution via Gumbel-Softmax [45].
- Using the masking mechanism to represent different candidate #channels and optimize its distribution via REINFORCE [50].

We indicate these strategies as “channel-wise interpolation”, “masking + Gumbel-Softmax”, and “masking + sampling” in Figure 8b. “channel-wise interpolation” can quickly find much better model than masking-based strategies. It might because that the interpolation strategy allows us to implicitly evaluate and compare two candidate #channels in each layer during each search step. In contrast, the masking strategies can only evaluate one candidate during each search step.

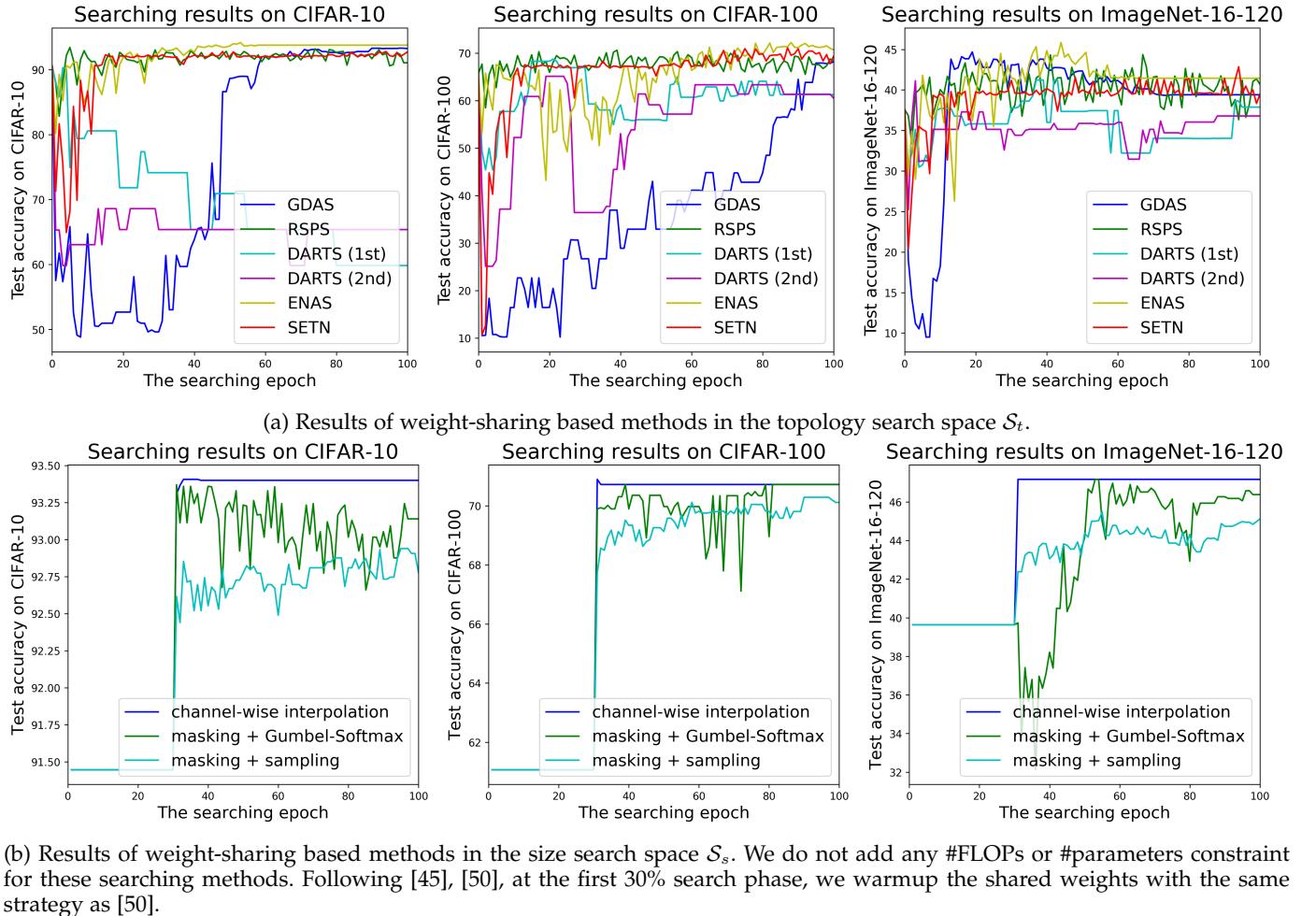
Since the original hyperparameters of [24], [45], [50] are designed for ImageNet and joint searching of filters and operations, they might be sub-optimal for the settings in NATS-Bench. In addition, we re-implement these

algorithms based on our codebase. They may have some differences compared to the original implementation due to the different search spaces, libraries, etc. Therefore, it is under investigation of whether our empirical observations can generalize to other scenarios or not.

### 5.3.3 Weight-sharing vs. Multi-trial based Methods

The weight-sharing based methods and multi-trial based methods have their unique advantages and disadvantages. Multi-trial based methods can theoretically find the best architecture as long as the proxy task is accurate, and the number of trials is large enough. However, their prohibitive computational cost has motivated researchers to design efficient weight-sharing based algorithms. However, sharing weights sacrifices the accuracy of each architecture candidate. As the search space increases, the shared weights are usually not able to distinguish the performance of different candidates.

**Clarification.** We have tried our best to implement each method using their reported best experimental set ups. However, please be aware that some algorithms might still result in sub-optimal performance since their hyperparameters might not be optimal for our NATS-Bench. We empirically found that some NAS algorithms are sensitive to some hyperparameters, and we have tried to compare



(b) Results of weight-sharing based methods in the size search space  $\mathcal{S}_s$ . We do not add any #FLOPs or #parameters constraint for these searching methods. Following [45], [50], at the first 30% search phase, we warmup the shared weights with the same strategy as [50].

Fig. 8: The test accuracy of the searched architecture candidate after each search epoch. We run different searching algorithms three times on three datasets. We plot the test accuracy of their searched model after each search epoch for the corresponding dataset. This test accuracy is evaluated after fully training the model on the corresponding dataset and averaged over three runs. We report more details including the accurate numbers and variance in Table 4.

them in as fair a way as possible. If researchers can provide better results with different hyperparameters, we are happy to update the benchmarks according to the new experimental results. We also welcome more NAS algorithms to be tested on our dataset and would be happy to include them accordingly.

## 6 DISCUSSION

**How to avoid over-fitting on NATS-Bench?** Our NATS-Bench provides a benchmark for NAS algorithms, aiming to provide a fair and computationally cost-friendly environment to the NAS community. The trained architecture and the easy-to-access performance of each architecture might provide some insidious ways for designing algorithms to over-fit the best architecture in our NATS-Bench. Thus, we propose some rules to follow in order to achieve the original intention of NATS-Bench, a fair and efficient benchmark.

1. *No regularization for a specific operation.* Since the best architecture is known in our benchmark, specific designs to fit the structural attributes of the best performing architecture constitute one of the insidious ways to fit our NATS-Bench. For example, as mentioned in Section 5, we found

that the best architecture with the same number of parameters for CIFAR-10 on NATS-Bench is ResNet. Restrictions on the number of residual connections is a way to over-fit the CIFAR-10 benchmark. While this can give a good result on this benchmark, the searching algorithm might not generalize to other benchmarks.

2. *Use the same meta hyperparameter for different datasets and search spaces in NATS-Bench.* The searching algorithm has some meta hyperparameter that controls the behaviour of search. For example, the temperature  $\tau$  in GDAS or the band width factor in BOHB. Using the same meta hyperparameter could evaluate the robustness of the searching algorithm and prevent it from over-fitting to a specific dataset.

3. *Use the provided performance.* The training strategy affects the performance of the architecture. We suggest to stick to the performance provided in our benchmark even if it is feasible to use other  $\mathcal{H}$  to get a better performance. This provides a fair comparison with other algorithms.

4. *Report results of multiple searching runs.* Since our benchmark can help to largely decrease the computational cost for a number of algorithms, multiple searching runs, which give stable results of the searching algorithm with acceptable

Search Space	Methods		CIFAR-10		CIFAR-100		ImageNet-16-120	
	Type	Name	validation	test	validation	test	validation	test
Topology Search Space $\mathcal{S}_t$	Multi-trial	REA	91.25±0.31	94.02±0.31	72.28±0.95	72.23±0.84	45.71±0.77	45.77±0.80
		REINFORCE	91.12±0.25	93.90±0.26	71.80±0.94	71.86±0.89	45.37±0.74	45.64±0.78
		RANDOM	91.07±0.26	93.86±0.23	71.46±0.97	71.55±0.97	45.03±0.91	45.28±0.97
		BOHB	91.17±0.27	93.94±0.28	72.04±0.93	72.00±0.86	45.55±0.79	45.70±0.86
	Weight Sharing	RSPS	87.60±0.61	91.05±0.66	68.27±0.72	68.26±0.96	39.73±0.34	40.69±0.36
		DARTS (1st)	49.27±13.44	59.84±7.84	61.08±4.37	61.26±4.43	38.07±2.90	37.88±2.91
		DARTS (2nd)	58.78±13.44	65.38±7.84	59.48±5.13	60.49±4.95	37.56±7.10	36.79±7.59
		GDAS	89.68±0.72	93.23±0.58	68.35±2.71	68.17±2.50	39.55±0.00	39.40±0.00
		SETN	90.00±0.97	92.72±0.73	69.19±1.42	69.36±1.72	39.77±0.33	39.51±0.33
		ENAS	90.20±0.00	93.76±0.00	70.21±0.71	70.67±0.62	40.78±0.00	41.44±0.00
	ResNet Optimal		90.86 91.61	93.91 94.37 (94.37)	70.50 73.49	70.89 73.51 (73.51)	44.10 46.73	44.23 46.20 (47.31)
Size Search Space $\mathcal{S}_s$	Multi-trial	REA	90.37±0.20	93.22±0.16	70.23±0.50	70.11±0.61	45.30±0.69	45.94±0.92
		REINFORCE	90.25±0.23	93.16±0.21	69.84±0.59	69.96±0.57	45.06±0.77	45.71±0.93
		RANDOM	90.10±0.26	93.03±0.25	69.57±0.57	69.72±0.61	45.01±0.74	45.42±0.86
		BOHB	90.07±0.28	93.01±0.24	69.75±0.60	69.90±0.60	45.11±0.69	45.56±0.81
	Weight Sharing	channel-wise interpolation masking + Gumbel-Softmax masking + sampling	90.71±0.00 90.41±0.10 89.73±0.37	93.40±0.00 93.14±0.13 92.78±0.30	70.30±0.00 70.30±0.00 69.67±0.22	70.72±0.00 70.72±0.00 70.11±0.33	44.73±0.00 45.71±0.39 44.70±0.60	47.17±0.00 46.38±0.27 45.11±0.76
	Largest Candidate Optimal		90.71 90.71	93.40 93.40 (93.65)	70.30 70.92	70.72 70.12 (71.34)	44.73 46.73	47.17 45.10 (47.40)

TABLE 4: We evaluate 13 different searching algorithms in our NATS-Bench. We use these algorithms to search for the architectures on different datasets and report *the accuracy of their discovered architectures* on the corresponding dataset. For multi-trial based methods, we run each algorithm 500 times and report the mean±variance. For weight-sharing based methods, we run each algorithm 3 times and report the mean±variance. “ResNet” indicates the candidate in topology search space with the same cell structure as the residual network [1]. “Largest Candidate” indicates the candidate in size search space with each #channels of 64. For “Optimal”, we report the performance of the candidate with the highest validation accuracy on each dataset, and we also include the test accuracy in the parentheses.

Notably, compared to the old version of this table in [28], we include the searching results on three (instead of one) datasets with better hyperparameters.

time cost, are strongly recommended.

**Limitation with regard to hyperparameter optimization (HPO).** The performance of an architecture depends on the hyperparameters  $\mathcal{H}$  for its training and the optimal configuration of  $\mathcal{H}$  may vary for different architectures. In NATS-Bench, we use the same configuration for all architectures, which may bring biases to the performance of some architectures. One related solution is HPO, which aims to search for the optimal hyperparameter configuration. However, searching for the optimal hyperparameter configurations and the architecture in one shot is too computationally expensive and still is an open problem [51].

**Potential extension of NATS-Bench.** Despite the straightforward extension by introducing HPO into NATS-Bench, there are some other interesting directions. One tendency in NAS is the cost constrained searching. For example, how to design a FLOPs constrain loss to regularize the discovered architecture to be efficient [24], [44], [45]? Since the latency and FLOPs information are off-the-shelf in NATS-Bench, our NATS-Bench can also be used to benchmark NAS algorithms using different kinds of cost loss.

**Potential designs using diagnostic information in NATS-Bench.** As pointed in Section 3.4, different kinds of diagnostic information are provided. We hope that more insights about NAS could be found by analyzing these diagnostic information and further motivate potential solutions for NAS. For example, parameter sharing [5] is the crucial technique to improve searching efficiency, but shared

parameter would sacrifice the accuracy of each architecture. Could we find a better way to share parameters of each architecture from the learned thousands of models’ parameters? Could we design new algorithms to take the mutual benefits of both multi-trial and weight sharing based methods?

**Generalization ability of the search space.** It is important to test the generalization capability of the empirical observations on this dataset. One possible strategy is to do all benchmark experiments on a much larger search space. Unfortunately, it is prohibitive regarding the expensive computational cost. We bring some results from [18], [20], [52] to provide some preliminary evidence of generalization. In Figure 7, we show the rankings of RANDOM, REA, and REINFORCE is (REA  $\geq$  REINFORCE  $\geq$  RANDOM). This is consistent with results in NAS-Bench-101, which contains more architecture candidates. For NAS methods with parameter sharing, we find that GDAS  $\geq$  DARTS (2nd)  $\geq$  DARTS (1st), which is also consistent with results in NAS-Bench-1SHOT1. Therefore, though it is not guaranteed, observations from our NATS-Bench have a potential to generalize to other search spaces.

**Acknowledgments:** We would like to thank Gabriel Bender, Pieter-Jan Kindermans, and Hanxiao Liu for their suggestions on experiments. Part of this project was supported by Google Cloud Credits from GCP Education Programs.

## REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
- [4] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *Proc. Int. Conf. Learn. Representations*, 2017.
- [5] H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean, "Efficient neural architecture search via parameters sharing," in *Proc. Int. Conf. Machine Learning*, 2018, pp. 4095–4104.
- [6] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 4780–4789.
- [7] X. Dong and Y. Yang, "Searching for a robust neural architecture in four gpu hours," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1761–1770.
- [8] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in *Proc. Int. Conf. Learn. Representations*, 2019.
- [9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [11] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8697–8710.
- [12] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2820–2828.
- [13] G. Ghiassi, T.-Y. Lin, and Q. V. Le, "Dropblock: A regularization method for convolutional networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 10727–10737.
- [14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [15] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Representations*, 2017.
- [16] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy, "Progressive neural architecture search," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 19–34.
- [17] X. Dong and Y. Yang, "One-shot neural architecture search via self-evaluated template network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3681–3690.
- [18] C. Ying, A. Klein, E. Real, E. Christiansen, K. Murphy, and F. Hutter, "NAS-Bench-101: Towards reproducible neural architecture search," in *Proc. Int. Conf. Machine Learning*, 2019, pp. 7105–7114.
- [19] A. Klein and F. Hutter, "Tabular benchmarks for joint architecture and hyperparameter optimization," *arXiv preprint arXiv:1905.04970*, 2019.
- [20] A. Zela, J. Siems, and F. Hutter, "NAS-BENCH-1SHOT1: Benchmarking and dissecting one shot neural architecture search," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [21] Y. Shu, W. Wang, and S. Cai, "Understanding architectures learnt by cell-based neural architecture search," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [22] Y.-C. Gu, Y. Liu, Y. Yang, Y.-H. Wu, S.-P. Lu, and M.-M. Cheng, "DOTS: Decoupling operation and topology in differentiable architecture search," *arXiv preprint arXiv:2010.00969*, 2020.
- [23] X. Dong, J. Huang, Y. Yang, and S. Yan, "More is less: A more complicated network with less inference complexity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5840–5848.
- [24] X. Dong and Y. Yang, "Network pruning via transformable architecture search," in *Proc. Advances Neural Inf. Process. Syst.*, 2019, pp. 760–771.
- [25] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *Proc. Int. Conf. Learn. Representations*, 2017.
- [26] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10781–10790.
- [27] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 82–92.
- [28] X. Dong and Y. Yang, "NAS-Bench-201: Extending the scope of reproducible neural architecture search," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [29] L. Li and A. Talwalkar, "Random search and reproducibility for neural architecture search," in *The Conf. on Uncertainty in Artificial Intelligence*, 2019.
- [30] K. Yu, C. Sciuco, M. Jaggi, C. Musat, and M. Salzmann, "Evaluating the search phase of neural architecture search," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [32] J. Yu and T. S. Huang, "Universally slimmable networks and improved training techniques," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1803–1811.
- [33] Y. He, J. Lin, Z. Liu, H. Wang, L.-J. Li, and S. Han, "AMC: Automl for model compression and acceleration on mobile devices," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 784–800.
- [34] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Citeseer, Tech. Rep.*, 2009.
- [35] P. Chrabaszcz, I. Loshchilov, and F. Hutter, "A downsampled variant of imagenet as an alternative to the cifar datasets," *arXiv preprint arXiv:1707.08819*, 2017.
- [36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
- [37] S. Falkner, A. Klein, and F. Hutter, "BOHB: Robust and efficient hyperparameter optimization at scale," in *Proc. Int. Conf. Machine Learning*, 2018, pp. 1436–1445.
- [38] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *The Journal of Machine Learning Research (JMLR)*, vol. 18, no. 1, pp. 6765–6816, 2018.
- [39] B. Baker, O. Gupta, R. Raskar, and N. Naik, "Accelerating neural architecture search using performance prediction," in *Proc. Int. Conf. Learn. Representations Workshop*, 2018.
- [40] H. Cai, T. Chen, W. Zhang, Y. Yu, and J. Wang, "Efficient architecture search by network transformation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 2787–2794.
- [41] T. Chen, I. Goodfellow, and J. Shlens, "Net2net: Accelerating learning via knowledge transfer," in *Proc. Int. Conf. Learn. Representations*, 2016.
- [42] C. Zhang, M. Ren, and R. Urtasun, "Graph hypernetworks for neural architecture search," in *Proc. Int. Conf. Learn. Representations*, 2019.
- [43] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "SMASH: one-shot model architecture search through hypernetworks," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [44] H. Cai, C. Gan, and S. Han, "Once for all: Train one network and specialize it for efficient deployment," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [45] A. Wan, X. Dai, P. Zhang, Z. He, Y. Tian, S. Xie, B. Wu, M. Yu, T. Xu, K. Chen *et al.*, "FBNetV2: Differentiable neural architecture search for spatial and channel dimensions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12965–12974.
- [46] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5927–5935.
- [47] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, pp. 239–251, 1945.
- [48] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *The Journal of Machine Learning Research (JMLR)*, vol. 13, no. Feb, pp. 281–305, 2012.
- [49] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3–4, pp. 229–256, 1992.
- [50] G. Bender, H. Liu, B. Chen, G. Chu, S. Cheng, P.-J. Kindermans, and Q. V. Le, "Can weight sharing outperform random archi-

- ture search? an investigation with tunas," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14 323–14 332.
- [51] X. Dong, M. Tan, A. W. Yu, D. Peng, B. Gabrys, and Q. V. Le, "AutoHAS: Efficient hyperparameter and architecture search," *arXiv preprint arXiv:2006.03656*, 2020.
- [52] D. Peng, X. Dong, E. Real, M. Tan, Y. Lu, G. Bender, H. Liu, A. Kraft, C. Liang, and Q. Le, "PyGlove: Symbolic programming for automated machine learning," in *Proc. Advances Neural Inf. Process. Syst.*, 2020.

## APPENDIX

### APPLICATION PROGRAMMING INTERFACE (API)

Users can easily query all information of an architecture by using our API, such as latency, training time, number of parameters, validation accuracy, etc. In this section, we show some example codes to query them.

Listing 1: Create an instance of our benchmark.

```
from nats_bench import create
# Load the data of the topology search space
nats_bench = create(search_space='topology')
# Load the data of the size search space
nats_bench = create(search_space='size')
```

Listing 2: Show the structure of each architecture.

```
amount = len(nats_bench)
for i, arch_str in enumerate(nats_bench):
    print('{:}/{:} : {}'.format(i, amount, arch_str))
```

Listing 3: Query the data of 115<sub>th</sub> architecture when training with 90 epochs ( $\mathcal{H}^1$ ); find the architecture with the highest accuracy on the validation set of CIFAR-100.

```
info = nats_bench.query_meta_info_by_index(
    arch_index=115, hp='90')
index, accuracy = nats_bench.find_best(
    dataset='cifar100', metric_on_set='valid')
```

Listing 4: Query the configuration of 12<sub>th</sub> architecture, its cost information, and its performance on different datasets.

```
config = nats_bench.get_net_config(
    arch_index=12, dataset='cifar10')
info = nats_bench.get_cost_info(
    arch_index=12, dataset='cifar10')
# The info is a dict, where key is train-loss,
# train-accuracy, train-all-time, test-loss, etc.
# The corresponding value is info[key].
info = nats_bench.get_more_info(
    arch_index=12, dataset='cifar10')
info = nats_bench.get_more_info(
    arch_index=12, dataset='cifar100')
info = nats_bench.get_more_info(
    arch_index=12, dataset='ImageNet16-120')
```

Listing 5: More advanced features.

```
# Query results of the 284-th architecture on
# CIFAR-100 when training with 12 epochs.
# The 'data' is a dict, where the key is the random
# seed and the value is the corresponding result.
data = nats_bench.query_by_index(
    arch_index=284, dataset='cifar100', hp='12')
# >> [777, 888, 999]
print(data.keys())
# Show the validation performance using the random
# seed of 888 for the 284-th architecture
info = results[888].get_eval('valid')
```

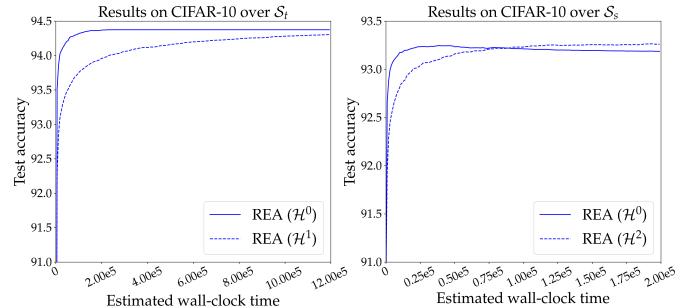


Fig. 9: The test accuracy of REA discovered architectures on CIFAR-10 over two search spaces.

### COMPARISON UNDER DIFFERENT TRAINING EPOCHS

Multi-trial based algorithms are computationally expensive because they need to traverse many trials, and each trial may cost hours. On the contrary, low fidelity approximation is computationally efficient but provides less accurate feedback to train the model, such as using fewer training epochs for each trial in Figure 7a and Figure 7b. To investigate the trade-off between efficiency and accuracy, we choose REA – the best-performing multi-trial based algorithm on our NATS-Bench, and show its performance when using 12 epochs for training and 200 (or 90) epochs for training in Figure 9.

For the topology search space, REA with  $\mathcal{H}^0$  (12 epochs) quickly converges after 1e5 seconds on CIFAR-10; however, REA with  $\mathcal{H}^1$  (200 epochs) is still worse than REA with  $\mathcal{H}^0$  after 1.2e6 seconds (about 333 hours). For the size search space, REA with  $\mathcal{H}^0$  (12 epochs) quickly converges after 2e4 seconds. REA with  $\mathcal{H}^2$  (90 epochs) is worse than REA with  $\mathcal{H}^0$  at the beginning, whereas it gradually outperforms REA with  $\mathcal{H}^0$  after about 8e4 seconds (about 22 hours). A similar phenomenon occurs on CIFAR-100 and ImageNet-16-120.



**Xuanyi Dong** received the B.E. degree in Computer Science and Technology from Beihang University (BUAA), Beijing, China, in 2016. He is currently a Ph.D. student at School of Computer Science, University of Technology Sydney (UTS), Australia. His research interests include automated deep learning and its application to real world problems. In 2019, he received the Google PhD Fellowship.



**Lu Liu** received her bachelor's degree from South China University of Technology (SCUT), Guangzhou, China, in 2017. She is currently pursuing the Ph.D. at AII, University of Technology Sydney (UTS), Australia. Her current research interests include deep learning, machine learning, few-shot learning and meta-learning.



**Katarzyna Musial** Katarzyna Musial received the M.Sc. degree in computer science from the Wrocław University of Science and Technology (WrUST), Poland, the M.Sc. degree in software engineering from the Blekinge Institute of Technology, Sweden, in 2006, and the Ph.D. from WrUST, in November 2009. In November 2009, she was appointed as a Senior Visiting Research Fellow with Bournemouth University (BU), where she has been a Lecturer in informatics, since 2010. In November 2011, she joined Kings as a Lecturer in computer science. In September 2015, she returned to Bournemouth University as a Principal Academic in Computing, where she was a member of the Data Science Institute. In September 2017, she joined as an Associate Professor in network science with the School of Software, University of Technology Sydney, where she is currently a member of the Advanced Analytics Institute. Her research interests include complex networked systems, analysis of their dynamics and its evolution, adaptive and predictive modeling of their structure and characteristics, as well as the adaptation mechanisms that exist within such systems are in the center of her research interests.



**Bogdan Gabrys** (SM'06) received the M.Sc. degree in electronics and telecommunication from Silesian Technical University, Gliwice, Poland, in 1994, and the Ph.D. degree in computer science from Nottingham Trent University, Nottingham, U.K., in 1998.

Over the last 25 years, he has been working at various universities and research and development departments of commercial institutions. He is currently a Professor of Data Science and a Director of the Advanced Analytics Institute at

the University of Technology Sydney, Sydney, Australia. His research activities have concentrated on the areas of data science, complex adaptive systems, computational intelligence, machine learning, predictive analytics, and their diverse applications. He has published over 180 research papers, chaired conferences, workshops, and special sessions, and been on program committees of a large number of international conferences with the data science, computational intelligence, machine learning, and data mining themes. He is also a Senior Member of the Institute of Electrical and Electronics Engineers (IEEE), a Member of IEEE Computational Intelligence Society and a Fellow of the Higher Education Academy (HEA) in the UK. He is frequently invited to give keynote and plenary talks at international conferences and lectures at internationally leading research centres and commercial research labs. More details can be found at: <http://bogdan-gabrys.com>