FairNAS: Rethinking Evaluation Fairness of Weight Sharing Neural Architecture Search

Xiangxiang Chu*  Bo Zhang  Ruijun Xu
{chuxiangxiang,zhangbo11,xuruijun}@xiaomi.com

Abstract

One of the most critical problems in weight-sharing neural architecture search is the evaluation of candidate models within a predefined search space. In practice, a one-shot supernet is trained to serve as an evaluator. A faithful ranking certainly leads to more accurate searching results. However, current methods are prone to making misjudgments. In this paper, we prove that their biased evaluation is due to inherent unfairness in the supernet training. In view of this, we propose two levels of constraints: expectation fairness and strict fairness. Particularly, strict fairness ensures equal optimization opportunities for all choice blocks throughout the training, which neither overestimates nor underestimates their capacity. We demonstrate that this is crucial for improving the confidence of models’ ranking. Incorporating the one-shot supernet trained under the proposed fairness constraints with a multi-objective evolutionary search algorithm, we obtain various state-of-the-art models, e.g., FairNAS-A attains 77.5% top-1 validation accuracy on ImageNet.

1. Introduction

The advent of neural architecture search (NAS) has brought deep learning into an era of automation [55]. Abundant efforts have been dedicated to searching within carefully designed search space [56, 37, 45, 31, 46]. Meanwhile, the evaluation of a network’s performance is an important building block for NAS. Conventional approaches evaluate an enormous amount of models based on resource-devouring training [56, 45]. Recent attention has been drawn to improve its efficiency via parameter sharing [2, 29, 35, 49].

Generally speaking, the weight-sharing approaches all involve training a supernet that incorporates many candidate subnetworks. They can be roughly classified into two categories: those who couple searching and training within one stage [35, 29, 4, 42, 49] and others who decouple them into two stages, where the trained supernet is treated as an evaluator for final searching [2, 1, 15, 32]. The supernet is a so-called one-shot model.

Despite being widely utilized due to searching efficiency, weight sharing approaches are roughly built on empirical experiments instead of solid theoretical ground. Several fundamental issues remain to be addressed. Namely, a) Why is there a large gap between the range of supernet predicted accuracies and that of “ground-truth” ones by stand-alone training from scratch [2, 1]? b) How to build a good evaluator that neither overestimates nor underestimates subnetworks? c) Why does the weight-sharing mechanism work, if under some conditions?

In this paper, we attempt to answer the above three questions for two-stage weight-sharing approaches. We present Fair Neural Architecture Search (FairNAS) in which we train the supernet under the proposed fairness constraints to improve evaluation confidence. Our analysis and experiments are conducted in a widely used search space as in [4, 49, 15, 42], as well as a cell-based search space from a common benchmark NAS-Bench-201 [13]. The contributions can be summarized as follows.

Firstly, we prove it is due to unfair bias that the supernet misjudges submodels’ performance, which is inevitable in current one-shot approaches [2, 1].

Secondly, we propose two levels of fairness constraints: Expectation Fairness (EF) and Strict Fairness (SF). They are enforced to alleviate supernet bias and to boost evaluation capability. Both outperform the existing unfair approaches while SF performs best with a ranking (τ) of 0.7412 on NAS-Bench-201.

Thirdly, we unveil the root cause of the validity of single-path supernet training under our fairness perspective. That is, different choice blocks are interchangeable during the supernet training as they learn similar feature maps, according to their high cosine similarity measure, see Figure 2.

Last but not the least, our fair single-path sampling is memory-friendly, and its GPU costs can also be linearly amortized to the number of target models, see Fig. 1. We then incorporate our supernet with an EA-based multi-objective searching framework, from which we obtain three state-of-the-art networks within a single proxyless run at a cost of 12 GPU days on ImageNet.

*This work was done when all the authors were at Xiaomi AI Lab.
2. Fairness Taxonomy of Weight-sharing NAS

2.1. Review of Biased Supernet

On the one hand, supernet training and searching for good models are nested. In ENAS [35], the sampling policy \(\pi(m, \theta)\) of an LSTM controller [17] and a sampled sub-network \(m\) are alternatively trained. The final models are sampled again by the trained policy \(\pi\), one who has the highest reward on a mini-batch of validation data is finally chosen. DARTS [29] combines the supernet training and searching within a bi-level optimization where each operation is associated with a coefficient denoting its importance. Both two methods treat all subnetworks unequally and introduce gradually increasing biases through optimization. Those who have better initial performance are more likely to be sampled or to maintain higher coefficients, resulting in a suboptimal or an even worse solution. For instance, architectures from DARTS usually contain an excessive number of skip connections [52, 26], which damage the outcome performance. Therefore, the prior-learning DARTS is biased as per skip connections, while a random approach doesn’t suffer [25]. DARTS overrated ‘bad’ models (jammed by skip connections), meantime many other good candidates are depreciated.

On the other hand, the rest one-shot methods consider the trained supernet as a confident proxy, which we also follow, to predict the real performance of all subnetworks [2, 1, 15]. We emphasize that a reliable proxy supernet should neither severely overestimate nor underestimate the ground-truth performance of any model. The next searching stage is decoupled from training and it can be implemented with random sampling, evolutionary algorithms, or reinforcement learning.

SMASH [2] invents a hyper network (referred to as HyperNet \(H\)) to generate the weights of a neural architecture by its binary encoding. This HyperNet resembles a typical supernet in that they can both produce weights for any architecture in the search space. At each step, a model is randomly sampled and trained based on the generated weights from \(H\), and in turn, it updates the weights of \(H\). For a set of randomly sampled models, a correlation between predicted validation errors and ground-truth exists, but it has a large discrepancy between the ranges, i.e., 40%-90% vs. 25%-30% on CIFAR-100 [24].

One-Shot [1] involves a dynamic dropout rate for the supernet, each time only a subset is optimized. Apart from its training difficulty, there is also an evident performance gap of submodels with inherited weights compared with their ground-truth, i.e., 30%-90% vs. 92%-94.5% on CIFAR-10 [24].
2.2. Evaluating Supernets by Ranking Ability

Regardless of how supernets are trained, what matters most is how well they predict the performance of candidate models. To this end, a recent work [41] evaluates weight-sharing supernets with Kendall Tau ($\tau$) metric [22]. It measures the relation between one-shot models and stand-alone trained ones. The range of $\tau$ is from -1 to 1, meaning the rankings are totally reversed or completely preserved, whereas 0 indicates no correlation at all. Surprisingly, most recent approaches behave incredibly poorly on this metric [32, 35, 41]. A method based on time-consuming incomplete training only reaches an average $\tau$ of 0.474 [54].

It measures the relation between one-shot models and stand-alone trained ones. The range of $\tau$ is from -1 to 1, meaning the rankings are totally reversed or completely preserved, whereas 0 indicates no correlation at all. Surprisingly, most recent approaches behave incredibly poorly on this metric [32, 35, 41]. A method based on time-consuming incomplete training only reaches an average $\tau$ of 0.474 [54].

Given the above biased supernets, we are motivated to revisit one-shot approaches to discover what might be the cause of the obvious range disparity, and, how much does unfair training affect their ranking ability. We ponder that if the supernet is trained free of bias, will it improve evaluation confidence and narrow the accuracy gap? Next, we start with a formal discussion of fairness.

2.3. Formal Formulation of Fairness

What kind of fairness can we think of? Will fairness help to improve supernet performance and ranking ability? First of all, to remove the training difference between a supernet and its submodels, we scheme an equality principle on training modality.

Definition 2.1. Equality Principle. Training a supernet satisfies the equality principle if and only if it is in the same way how a submodel is trained.

Only those who train a single path model at each step meet this principle by its definition. On the contrary, other methods like DARTS [29] train the supernet with all paths altogether, One-Shot [1] dynamically drops out some paths, and ProxylessNAS [4] uses two paths, directly violating the principle.

Formally, we discuss fairness in a common supernet that consists of $L$ layers, each with several choice blocks. Without loss of generality, we suppose each layer has an equal number of choices, say $m$. A model is generated by sampling a block layer by layer. The weights are updated for $n$ times in total. Therefore, we can describe the training process as $P(m, n, L)$.

2.3.1 First Attempt: Expectation Fairness

In order to reduce the above mentioned bias in Section 2.1, a natural way is to guarantee all choices blocks have equal expectations after $n$ steps. We define this basic requirement as expectation fairness in Definition 2.2.

Definition 2.2. Expectation Fairness. On the basis of Definition 2.1, let $\Omega$ be the sampling space containing $m$ basic events $\{l_1, l_2, ..., l_m\}$, which are generated by selecting a block from layer $l$ with $m$ choice blocks. Let $Y_{l_i}$ be the number of times that the outcome $l_i$ is observed (updated) over $n$ trials. Then the expectation fairness is that for $P(m, n, L)$, $E(Y_{l_1}) = E(Y_{l_2}) = ... = E(Y_{l_m})$ holds, $\forall l \in L$.

2.3.2 An EF Example: Uniform Sampling

Let us check a single-path routine [15] which uses uniform sampling. As sampling on any layer $l$ is independent of others, we first consider the case $P(m, n, l)$. Selecting a block from layer $l$ is subject to a categorical distribution. In this case, each basic event occurs with an equal probability $p(X = l_i) = \frac{1}{m}$. For $n$ steps, the expectation and variance of $Y_{l_i}$ can be written as,

$$E(Y_{l_i}) = n \times p_{l_i} = n/m$$

$$\text{Var}(Y_{l_i}) = n \times p_{l_i} (1 - p_{l_i}) = \frac{n(m-1)}{m^2}$$

That’s to say, all choices share the same expectation and variance. Consequently, uniform sampling meets Expectation Fairness by Definition 2.2 and it seems superficially fair for various choices. However, Expectation Fairness is not enough. For example, we can randomly sample each model and keep it training for $k$ times, then switch to another. This procedure also meets Definition 2.2, but it’s very unstable to train.

Even in SPOS [15] with uniform sampling, there is a latent ordering issue. For a sequence of choices $(M_1, M_2, M_3)$, it implies an inherent training order $M_1 \rightarrow M_2 \rightarrow M_3$. Since each model is usually trained by back-propagation, the trained weights of $M_1$ are immediately updated to the supernet and those of $M_2$ are renewed next while carrying the effect of the former update, so for $M_3$. A simple permutation of $(M_1, M_2, M_3)$ does comply with Expectation Fairness but yields different results. Besides, if the learning rate $lr$ is changed within the sequence, the situation thus becomes even more complicated.

Generally, for $P(m, n, L)$ where $m$, $n$, $L$ are positive integers, assume the sampling times $n$ can be divided by $m(\geq 2)$. If we adopt uniform sampling, as $n$ goes infinite, it is impossible for $m$ choices to be sampled for an exactly equal number of times. This is formally stated as below.

Lemma 2.1. Regarding $P(m, n, L)$, $\forall n \in \{x : \% m = 0, x \in N_+\}$, $\lim_{n \rightarrow +\infty} p(Y_{l_1} = Y_{l_2} = ... = Y_{l_m}) = 0$.

Proof. Let $f(m, n) = p(Y_{l_1} = Y_{l_2} = ... = Y_{l_m})$.

$$f(m, n) = C_m^n \times C_m^{n-1} \times C_m^{n-2} \times ... \times C_m^1 \frac{1}{m^m} = \frac{n!}{(\frac{n}{m})^m} \frac{1}{m^n}$$ (2)

Firstly, we prove the existence of limitation, $f(n)$ strictly decreases monotonically with $n$ and $f(n) \geq 0$, therefore, its limitation exists.
Secondly, we calculate its limitation using equivalent infinity replacement based on Stirling’s approximation about factorial [48].

\[
\lim_{n \to +\infty} f(m, n) = \lim_{n \to +\infty} \frac{n!}{(\frac{n}{m})^m \times m^n} = \lim_{n \to +\infty} \frac{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n}{\sqrt{2\pi m} \left(\frac{m}{e}\right)^m}
\]

(3)

Q.E.D.

Lemma 2.1 is somewhat counter-intuitive and thereby neglected in previous works. To throw light on this phenomenon, we plot this probability curve in Figure 3. We see that \(f(2, n)\) decreases below 0.2 when \(n \geq 20\). In most cases, \(n \geq 10^6\), which suffers severely from this issue.

2.3.3 A Meticulous Overhaul: Strict Fairness

Our insights come from the above overlooked phenomenon. We propose a more rigorous requirement that ensures the parameter of every choice block be updated the same amount of times at any stage, which is called strict fairness and formally as Definition 2.3.

**Definition 2.3. Strict Fairness.** Regarding \(P(m, n, L)\), \(\forall n \in \{x : x|m = 0, x \in N^+\}\), \(Y_{l_1} = Y_{l_2} = \ldots = Y_{l_m}\) holds.

Definition 2.3 imposes a constraint more demanding than Definition 2.2. That is, \(p(Y_{l_1} = Y_{l_2} = \ldots = Y_{l_m}) = 1\) holds at any time. It seems subtle but it will be later proved to be crucial. Nevertheless, we have to be aware that it is not ultimate fairness since different models have their own optimal initialization strategy and hyperparameters, which we single them out for simplicity.

3. Fair Neural Architecture Search

Our NAS pipeline is divided into two stages: training the supernet and searching for competitive models.

3.1. Stage 1: Train Supernet with Strict Fairness

We first propose a fair sampling and training algorithm to strictly abide by Definition 2.3. We use *uniform sampling without replacement* and sample \(m\) models at step \(l\) so that each choice block must be activated and updated only once. This is detailed in Algorithm 1 and depicted by Figure 4.

![Figure 3. The function curve of \(f(m, n)\) in Lemma 2.1.](image)

**Algorithm 1 : Stage 1 - Fair Supernet Training.**

**Input:** training steps \(n\), search space \(S_{(m,L)}\), \(m \times L\) supernet parameters \(\Theta(m, L)\), search layer depth \(L\), choice blocks \(m\) per layer, training epochs \(N\), training data loader \(D\), loss function \(Loss\)

**initialize** every \(\theta\) in \(\Theta(m, L)\) do

for \(i = 1\) to \(N\) do

for data, labels in \(D\) do

for \(l = 1\) to \(L\) do

\(c_t = \text{an uniform index permutation for the choices of layer } l\)

end for

Clear gradients recorder for all parameters

\(\nabla \theta_{j,l} = 0, j = 1, 2, \ldots, m, l = 1, 2, \ldots, L\)

for \(k = 1\) to \(m\) do

Build \(model_k = (c_1_k, c_2_k, \ldots, c_k_L)\) from sampled index

Calculate gradients for \(model_k\) based on \(Loss\), data, labels.

Accumulate gradients for activated parameters,

\(\nabla \theta_{c_1_k,1}, \nabla \theta_{c_2_k,2}, \ldots, \nabla \theta_{c_k_L,L}\)

end for

update \(\theta_{(m,L)}\) by accumulated gradients.

end for

end for

end for

end for

To reduce the bias from different training orders, we don’t perform back-propagation and update parameters immediately for each model as in the previous works [1, 15]. Instead, we define one *supernet step* as several back-propagation operations (BPs) accompanied by a single parameter update. In particular, given a mini-batch of training data, each of \(m\) single-path models is trained with back-propagation. Gradients are then accumulated across the selected \(m\) models but supernet’s parameters get updated only when all \(m\) BPs are done. This approach also doesn’t suffer from the *ordering issue* as each choice block is updated regardless of external learning rate strategies.

**Strict Fairness Analysis.** We now check whether our proposed Algorithm 1 satisfies Strict Fairness. By its design, each choice block is activated only once during a parameter update step. Thus \(Y_{l_1} = Y_{l_2} = \ldots = Y_{l_m}\) holds. In particular, \(Y_{l_1} = Y_{l_2} = \ldots = Y_{l_m} = n/m\) holds. Here, we write its
expectation and variance as follows:

\[ E(Y_i^t) = n/m, \ Var(Y_i^t) = 0 \]  

(4)

Compared with Equation 1, the obvious difference lies in the variance. For the single-path approach with uniform sampling [15], the variance spreads along with \( n \), which gradually increases the bias. However, our approach calibrates this inclination and assures fairness at every step.

3.2. Stage 2: Searching with Supernet

For searching, we can either choose random search, vanilla evolutionary algorithms, or reinforcement learning. In practice, there are many requirements and objectives to achieve, e.g., inference time, multiply-adds, and memory costs, etc. This leads us to adopt a multi-objective solution. Besides, as the search space is too vast to enumerate all models, we need an efficient approach to balance the exploration and exploitation trade-off instead of a random sampling strategy. Here we adopt a searching algorithm from an NSGA-II [9] with a small variation by using Proximal Policy Optimization [40] as the default reinforcing algorithm. The whole process is given in Algorithm 2. Benefiting from the fast evaluation of the weight-sharing supernet, we can achieve tremendous speed-up in terms of GPU days by two orders of magnitudes.

4. Experiments

4.1. Setup

Search Space. We use several search spaces in this paper. The first one is that of NAS-Bench-201 [13], which is a new benchmark for NAS methods. Besides, We adopt two extra search spaces to compare with other NAS methods on ImageNet. (a) A search space for fairness analysis, which is based on MobileNetV2’s inverted bottleneck blocks as done in [4]. In particular, we retain the same amount of layers

\[ (6^{19} \times 5^{19} \times 4^{19} \times 3^{19} \times 2^{19} \times 1)/6! = (6!)^{18}. \]

3It can be calculated by \((6^{19} \times 5^{19} \times 4^{19} \times 3^{19} \times 2^{19} \times 1)/6! = (6!)^{18}. \]

Algorithm 2: Stage 2 - Search Strategy.

\textbf{Input}: Supernet \( SN \), the number of generations \( G \), validation dataset \( D \)

\textbf{Output}: A set of \( K \) individuals on the Pareto front.

Train supernet \( SN \) with Algorithm 1.

Uniform initialization for the populations \( P_i \) and \( Q_1 \).

for \( i = 1 \) to \( G \) do

\( R_i = P_i \cup Q_i \) \n
for all \( p \in R_i \) do

Evaluate model \( p \) with inherited weights from \( SN \) on \( D \)

end for

\( F = \text{non-dominated-sorting}(R_i) \)

Pick \( N \) individuals to form \( P_{i+1} \) by ranks and the crowding distance.

\( M = \text{tournament-selection}(P_{i+1}) \)

\( Q_{i+1} = \text{crossover}(M) \cup \text{hierarchical-mutation}(M) \)

end for

Select \( K \) evenly-spaced models from \( P_{G+1} \) to train

with standard MobileNetV2 [39]. Convolution kernels are with the size in \((3, 5, 7)\) and expansion rates are of \((3, 6)\). We keep the number of filters unchanged. Besides, the squeeze-and-excitation block [20] is excluded. In total, it has a size of \(6^{19}\). (b) A search space of 19 layers as ProxylessNAS [4], whose size spreads to \(6^{19}\). This is to be on par with various state-of-the-art methods.

Training Hyperparameters. For NAS-Bench-201, we train the supernet for 50 epochs using a batch size of 128. The initial learning rate is 0.025 and decayed to zero by the cosine schedule.

For search space (a), we train the supernet for 150 epochs by using a batch size of 256 and adopt a stochastic gradient descent optimizer with a momentum of 0.9 [43] based on standard data augmentation as [39]. A cosine learning rate decay strategy [30] is applied with an initial learning rate of 0.045. Moreover, We regularize the training with L2 weight
decay ($4 \times 10^{-5}$). Our supernet is thus trained to fullness in 10 GPU days.

For search space (b), we follow the same strategy as above for training the supernet, but we adopt vanilla data processing as well as training tricks in [45] for stand-alone models. Regarding the stand-alone training of sampled models, we use similar training tricks. To be consistent with the previous works, we don’t employ tricks like cutout [10] or mixup [53], although they can further improve the scores on the test set.

4.2. Search Result

4.2.1 Search on ImageNet.

Figure 8 (supplementary) exhibits the resulting FairNAS-A, B and C models, which are sampled from our Pareto front to meet different hardware constraints. The quantitative result is shown in Table 1.

Notably, FairNAS-A obtains a highly competitive result 75.3% top-1 accuracy for ImageNet classification, which surpasses MnasNet-92 (+0.5%) and Single-Path-NAS (+0.3%). FairNAS-B matches Proxyless-GPU with much fewer parameters and multiply-adds. Besides, it surpasses Proxyless-R Mobile (+0.5%) with a comparable amount of multiply-adds.

Our models also reach a new state of the art when equipped with combined tricks such as Squeeze-and-Excitation [20], Swish activations [36] and AutoAugment [8]. Namely, FairNAS-A obtains 77.5% top-1 accuracy by using similar FLOPS as EfficientNet-B0. Even without mixed kernels [47], FairNAS-B (77.2%) outperforms MixNet-M (77.0%) with 11M fewer FLOPS. The most light-weight model FairNAS-C (76.7%) also outperforms EfficientNet-B0 with about 20% fewer FLOPS.

4.2.2 Search on NAS-Bench-201.

To be comparable with existing methods, we formulate our problem as a single objective one: finding the best model. Specially, we use a standard evolutionary algorithm in the problem as a single objective one: finding the best model. For search space (b), we follow the same strategy as above for training the supernet, but we adopt vanilla data processing as well as training tricks in [45] for stand-alone models. Regarding the stand-alone training of sampled models, we use similar training tricks.

4.3. Transferred Results on CIFAR

To validate transferability of FairNAS models, we adapt the pre-trained models on ImageNet to CIFAR-10 and CIFAR-100 following the configuration of GPipe [21] and [23]. Table 3 shows that FairNAS models outperform the rest transferred models with higher top-1 accuracy.

4.4. Transferability on Object Detection

For object detection, we treat FairNAS models as drop-in replacements for RetinaNet’s backbone [27]. We follow the same setting as [27] and exploit MMDetection toolbox [5] for training. All the models are trained and evaluated on MS COCO dataset (train2017 and val2017 respectively) [28] for 12 epochs with a batch size of 16. The initial learning rate is 0.01 and decayed by 0.1 × at epochs 8 and 11.

The input features from these backbones to the FPN module are from the last depthwise layers of stage 2 to 5. The number of output channels of FPN is kept 256 as [27]. We also use $\alpha = 0.25$ and $\gamma = 2.0$ for the focal loss. Given longer training epochs and other tricks, the detection performance can be improved further. However, it’s sufficient to compare the transferability of various methods. The results are given in Table 4, we have the best transferability.

4.5. Transferability on Semantic Segmentation

We further evaluate FairNAS models as a feature extractor with DeepLabv3+[6] on the mobile semantic segmentation task, which confirms FairNAS backbones are competitive. All models are first pre-trained on COCO dataset [28], then coarsely trained on VOC2012 [14] extra annotated images.

---

Table 1. Comparison of mobile models on ImageNet. $M$: Memory cost at all sampled sub-models (SM), a single path or two paths (SP/TP), and a whole supernet (SN). : w/ SE and Swish, P: Number of parameters, ×: Cost shared among A, B and C, †: reportedly 3× EfficientNet Training.

| Models                  | ×+ (M) | $P$ (M) | Top-1 (%) | $M$ (GPU days) | Cost |
|-------------------------|--------|---------|-----------|----------------|------|
| MobileNetV2 [39]        | 300    | 3.4     | 72.0      | -              | -    |
| NASNet-A [56]           | 564    | 5.3     | 74.0      | SM             | 1800 |
| MnasNet-92 [45]         | 388    | 3.9     | 74.8      | SM             | ≈4k  |
| DARTS [29]              | 574    | 4.7     | 73.3      | SN             | 0.5  |
| PC-DARTS (CIFAR10) [50] | 586    | 5.3     | 74.9      | SN             | 0.1  |
| One-Shot Small (F=32) [1] | -    | 5.1     | 74.2      | SN             | 4    |
| AtomNAS-A [33]          | 258    | 3.9     | 74.6      | SN             | 20.5 |
| FBNet-B [49]            | 295    | 4.5     | 74.1      | TP             | 0.9  |
| Proxyless-R [4]         | 320    | 4.0     | 74.6      | TP             | 0.01 |
| Proxyless GPU [4]       | 465†   | 7.1     | 75.1      | TP             | 8.3  |
| Single Path One-Shot [15] | 323   | 3.5     | 74.4      | SP             | 12   |
| Single-Path NAS [42]    | 365    | 4.3     | 75.0      | SP             | 1.25 |
| **FairNAS-A** (Ours)    | 388    | 4.6     | 75.3      | SP             | 12   |
| FairNAS-B (Ours)        | 345    | 4.5     | 75.1      | SP             | 12   |
| FairNAS-C (Ours)        | 321    | 4.4     | 74.7      | SP             | 12   |
| MnasNet-A2 [45]         | 340    | 4.8     | 75.6      | SM             | ≈4k  |
| MobileNetV3 Large [18]  | 219    | 5.4     | 75.2      | SM             | ≈3k  |
| GhostNet 1.3 × [16]     | 226    | 7.3     | 75.7      | -              | -    |
| EfficientNet B0 [46]    | 390    | 5.3     | 76.3      | SM             | ≈3k  |
| OFA w/ PS #75 [3]       | 230    | 7.6     | 76.9      | SP             | 175  |
| BigNAS-S [51]           | 242    | 4.5     | 76.5      | SP             | 48   |
| MixNet-M [47]           | 360    | 5.0     | 77.0      | SM             | ≈3k  |
| AtomNAS-C+ [33]         | 363    | 5.9     | 77.6      | SN             | 20.5 |
| **FairNAS-A**† (Ours)   | 392    | 5.9     | 77.5      | SP             | 12   |
| FairNAS-B† (Ours)       | 349    | 5.7     | 77.2      | SP             | 12   |
| FairNAS-C† (Ours)       | 325    | 5.6     | 76.7      | SP             | 12   |

---

4.2. Search Result

4.2.1 Search on ImageNet.

Figure 8 (supplementary) exhibits the resulting FairNAS-A, B and C models, which are sampled from our Pareto front to meet different hardware constraints. The quantitative result is shown in Table 1.

Notably, FairNAS-A obtains a highly competitive result 75.3% top-1 accuracy for ImageNet classification, which surpasses MnasNet-92 (+0.5%) and Single-Path-NAS (+0.3%). FairNAS-B matches Proxyless-GPU with much fewer parameters and multiply-adds. Besides, it surpasses Proxyless-R Mobile (+0.5%) with a comparable amount of multiply-adds.

Our models also reach a new state of the art when equipped with combined tricks such as Squeeze-and-Excitation [20], Swish activations [36] and AutoAugment [8]. Namely, FairNAS-A obtains 77.5% top-1 accuracy by using similar FLOPS as EfficientNet-B0. Even without mixed kernels [47], FairNAS-B (77.2%) outperforms MixNet-M (77.0%) with 11M fewer FLOPS. The most light-weight model FairNAS-C (76.7%) also outperforms EfficientNet-B0 with about 20% fewer FLOPS.

4.2.2 Search on NAS-Bench-201.

To be comparable with existing methods, we formulate our problem as a single objective one: finding the best model. Specially, we use a standard evolutionary algorithm in the second stage after the supernet is fairly trained. The result is shown in Table 2. Our method outperforms the other baselines in most datasets with the lowest search cost.

4.3. Transferred Results on CIFAR

To validate transferability of FairNAS models, we adapt the pre-trained models on ImageNet to CIFAR-10 and CIFAR-100 following the configuration of GPipe [21] and [23]. Table 3 shows that FairNAS models outperform the rest transferred models with higher top-1 accuracy.

4.4. Transferability on Object Detection

For object detection, we treat FairNAS models as drop-in replacements for RetinaNet’s backbone [27]. We follow the same setting as [27] and exploit MMDetection toolbox [5] for training. All the models are trained and evaluated on MS COCO dataset (train2017 and val2017 respectively) [28] for 12 epochs with a batch size of 16. The initial learning rate is 0.01 and decayed by 0.1 × at epochs 8 and 11.

The input features from these backbones to the FPN module are from the last depthwise layers of stage 2 to 5. The number of output channels of FPN is kept 256 as [27]. We also use $\alpha = 0.25$ and $\gamma = 2.0$ for the focal loss. Given longer training epochs and other tricks, the detection performance can be improved further. However, it’s sufficient to compare the transferability of various methods. The results are given in Table 4, we have the best transferability.

4.5. Transferability on Semantic Segmentation

We further evaluate FairNAS models as a feature extractor with DeepLabv3+[6] on the mobile semantic segmentation task, which confirms FairNAS backbones are competitive. All models are first pre-trained on COCO dataset [28], then coarsely trained on VOC2012 [14] extra annotated images.
## 5. Ablation Study

### 5.1. Model Ranking Capacity

As stated, the most important role of the supernet in the two-stage methods is to score models’ relative performance, i.e. model ranking.

For supernet training, we set up three control groups that meet Expectation Fairness as our baselines. a) EF $lr$, uniformly sampling one path and train $k$ times, followed by parameter update. b) EF $1/lr$: same as the first one except that the learning rate is scaled by $\frac{1}{k}$. In practice, we set $k = 6$ to make it comparable to FairNAS. c) SPOS: an implementation of Single-Path One-Shot [15]. Other hyperparameters are kept the same. Note a), c) and FairNAS all use the same learning rate $lr$.

We run the search pipeline for 200 epochs with a population size of 64, sampling 12,800 models in total. It takes only 2 GPU days due to accelerated evaluation. Due to high training cost, we sampled 13 models at approximately equal distances on the Pareto front and trained them from scratch to get the ranking, which is shown to the right of Figure 1. We observe that the FairNAS supernet gives a highly relevant ranking while Single-Path One-Shot [15] doesn’t. The training process of sampled models is plotted in Figure 9 (see supplementary).

We further adopt Kendall Tau [22] for the ranking analysis following a recent work [41] that evaluates NAS approaches. A method based on incomplete training reaches an average $\tau$ of 0.474 [54]. Instead, we hit a new high record of the Kendall rank correlation coefficient $\tau = 0.9487$. We show our ranking comparison with baseline groups in Table 6. In general, methods with EF have a better ranking than those without EF, while SF is the best of all, which discloses the relevance of fairness to ranking in one-shot approaches.

### Table 2. Comparison on NAS-Bench-201 [13]. Averaged on three runs of searching

| Method       | Cost (seconds) | CIFAR-10 | CIFAR-100 | ImageNet16-120 |
|--------------|----------------|----------|-----------|----------------|
| DARTS [29]   | 11625          | 39.77±0.00 | 54.30±0.00 | 15.03±0.00 |
| ENAS [35]    | 14058          | 37.51±3.19 | 53.89±0.58 | 13.37±2.35 |
| SETN [11]    | 34139          | 84.04±0.28 | 87.64±0.00 | 58.86±0.06 |
| GDAS [12]    | 31609          | 89.89±0.08 | 93.61±0.09 | 71.34±0.04 |
| FairNAS (ours) | 9845         | 90.07±0.57 | 93.23±0.18 | 70.94±0.94 |

| Models | Input Size | CIFAR-10 | CIFAR-100 | ImageNet16-120 |
|--------|------------|----------|-----------|----------------|
| NASNet-A Large [56] | 331×331 | 12030 98.0 | 12031 98.7* | 75.32 360 |
| EfficientNet-B0 [46] | 224×224 | 387 98.1 | 387 98.6* | 77.5 224 |
| MixNet-M [47] | 224×224 | 359 97.9 | 359 97.1* | 77.10 375 |
| FairNAS-A | 224×224 | 391 98.2 | 391 98.7 | 87.3 224 |
| FairNAS-B | 224×224 | 348 98.1 | 348 98.7 | 87.0 224 |
| FairNAS-C | 224×224 | 324 98.0 | 324 98.6 | 86.7 224 |

| Methods | Fairness | $\tau_a$ | $\tau_N$ |
|---------|----------|---------|---------|
| One-Shot [1]† | None | 0.1245 | 0.0934 |
| Uniform ($k = 6$, baseline) | EF | 0.4871 | 0.3651 |
| Uniform ($k = 1$, 1/lr) | EF | 0.4871 | 0.4072 |
| SPOS [15] (k = 1)† | EF | 0.6153 | 0.5681 |
| FairNAS† | SF | 0.9487 | 0.7412 |

| Backbones | × + Acc (M) (%) | AP (M) (%) | AP50 (M) (%) | AP75 (M) (%) | APs (M) (%) | APM (M) (%) | APx (M) (%) |
|-----------|-----------------|------------|-------------|-------------|-------------|-------------|-------------|
| MobileNetV2 [39] | 300 72.0 | 28.3 | 14.8 | 30.7 | 38.1 |
| SingPath NAS [42] | 365 75.0 | 30.7 | 32.2 | 33.9 | 41.6 |
| MobileNetV3 [18] | 219 75.2 | 29.9 | 30.8 | 33.3 | 41.1 |
| MnasNet-A2 [45] | 340 75.6 | 30.5 | 32.0 | 34.1 | 41.1 |
| MixNet-M [47] | 360 77.0 | 31.3 | 31.7 | 33.5 | 41.9 |
| FairNAS-A | 392 77.5 | 32.4 | 33.9 | 36.3 | 43.2 |
| FairNAS-B | 349 77.2 | 31.7 | 33.0 | 35.2 | 42.5 |
| FairNAS-C | 325 76.7 | 31.2 | 32.7 | 34.4 | 42.3 |

| Network | OS | ASPP | Params (M) | × + | mIOU |
|---------|----|------|------------|-----|------|
| MobileNetV1 [19] | 16 | ✓ | 11.15 | 14.25B | 75.29 |
| MobileNetV2 [39] | 16 | X | 4.52M | 2.75B | 75.32 |
| FairNAS-A | 16 | X | 3.26M | 3.98B | 78.54 |
| FairNAS-B | 16 | X | 3.11M | 3.74B | 77.10 |
| FairNAS-C | 16 | X | 3.01M | 3.60B | 77.64 |

Table 5. Semantic Segmentation on VOC 2012. OS: Output Stride
5.2. Comparisons of Searching Algorithms

For the second-stage, we adopt multi-objective optimization where three objectives are considered: accuracies, multiply-adds, and the number of parameters. Specifically, we apply MoreMNAS with a minor modification in which PPO [40] is utilized instead of REINFORCE [44].

We construct several comparison groups that cover the main searching algorithms: a) EA: NSGA-II with reinforced mutation, b) random search, c) Multi-objective RL: MnasNet which uses PPO with a mixed multi-objective reward [45]. The results are shown in Figure 5, control groups generally align within our Pareto front and are constricted within a narrow range, affirming an excellent advantage in the MoreMNAS variant.

![Figure 5. Pareto front (last generation elitists $P_{G+1}$) of the MoreMNAS variant (adopted) compared with Left: NSGA2 (EA-like baseline) with RL mutator and random search (random baseline), Right: MnasNet (RL baseline). Each samples 1,088 models.](image)

5.3. Component Contribution Analysis

Being a two-stage method, which stage contributes more to the final performance of the architecture? The experiment in NAS-Bench-201 has answered this question. To be complete, we further compare various supernet training strategies while fixing the second stage using the ImageNet dataset. Considering it’s not affordable to train the entire models from Pareto front, we impose an explicit constraint of maximum 400M FLOPS. The best models from One-Shot [2] (no EF) and SPOS (EF) [7] reach 74.0% and 74.6% top-1 accuracies, indicating that our result (75.3%) benefits mainly from the ranking capacity of the supernet in the first stage.

6. Discussions

6.1. Why Does Single-Path Training Work?

Our supernet generates a relatively small range of one-shot accuracies, from which we postulate that choice blocks be quite alike in terms of capacity. In fact, given an input of a chickadee image, the choice blocks of the first layer yield similar feature maps on the same channel, as shown in Figure 2. But how much do they resemble each other? We involve the cosine similarity [34] to measure the distance among various feature vectors. It ranges from -1 (opposite) to 1 (identical), where 0 indicates no correlation. In Figure 6, each 6 × 6 symmetric matrix shows the cross-block distances per channel, they are very similar (above 0.9).

![Figure 6. Cross-block channel-wise cosine similarity matrix on feature maps of 6 choice blocks in Layer 1. We observe that each choice block learns very similar features on the same channel.](image)

In summary, the channel-wise feature maps generated by our supernet come with high similarities. We conclude that this important characteristic significantly stabilizes the whole training process. For layer $l + 1$, its input is randomly from choice blocks in the previous layer $l$. As different choices have highly similar channel-aligned features, the random sampling constructs a mechanism mimicking feature augmentation, which boosts the supernet training.

6.2. Fairness Closes Supernet Accuracy gap

As discussed in Section 2.1, previous one-shot methods [2, 1] have a large accuracy gap between the one-shot and stand-alone models. We define it as supernet accuracy gap, $\lambda = |\delta_{oneshot} - \delta_{standalone}|$, where $\delta_{oneshot}$ is the accuracy range of one-shot models, and $\delta_{standalone}$ for stand-alone models. Ideally, $\delta_{oneshot}$ can be obtained by evaluating all paths from the supernet but not affordable since the search space is enormous. Instead, we can approximate $\delta_{oneshot}$ by covering a wide range of models. We randomly sample 1,000 models from our supernet, then we evaluate these models directly on the ImageNet validation set. Their top-1 accuracies (see Figure 1) range from 0.666 to 0.696, which leads to $\delta_{oneshot} = 0.03$, hence it reduces $\lambda$ as well.

7. Conclusion

In this work, we scrutinize the weight-sharing neural architecture search with a fairness perspective. Observing that unfairness inevitably incurs a severely biased evaluation of one-shot model performance, we propose two degrees of fairness enhancement, where Strict Fairness (SF) works best. Our supernet trained under SF then acts as a performance evaluator. In principle, the fair supernet can be incorporated in any search pipeline that requires an evaluator. To demonstrate its effectiveness, we adopt a multi-objective evolutionary backend. After searching proxylessly on ImageNet for 12 GPU days, we harvest three state-of-the-art models of different magnitudes nearby Pareto Optimality. Future works remain as to study fairness under heterogenous search spaces and to improve the evaluation performance of the supernet.
References

[1] Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, and Quoc Le. Understanding and Simplifying One-Shot Architecture Search. In International Conference on Machine Learning, pages 549–558, 2018. 1, 2, 3, 4, 6, 7, 8, 11

[2] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. SMASH: One-Shot Model Architecture Search through HyperNetworks. In International Conference on Learning Representations, 2018. 1, 2, 8, 11

[3] Han Cai, Chuang Gan, and Song Han. Once for All: Train One Network and Specialize it for Efficient Deployment. In International Conference on Learning Representations, 2020. 6

[4] Han Cai, Ligeng Zhu, and Song Han. ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. In International Conference on Learning Representations, 2019. 1, 3, 5, 6, 11

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155, 2019. 6, 13

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In European Conference on Computer Vision, 2018. 6

[7] Xiangxiang Chu, Bo Zhang, and Ruijun Xu. Multi-Objective Reinforced Evolution in Mobile Neural Architecture Search. In European Conference on Computer Vision Workshop, 2020. 11, 12

[8] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. AutoAugment: Learning Augmentation Policies from Data. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6

[9] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A Fast and Elitist Multi-objective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 6(2):182–197, 2002. 5

[10] Terrance DeVries and Graham W Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. arXiv preprint arXiv:1708.04552, 2017. 6

[11] Xuanyi Dong and Yi Yang. One-Shot Neural Architecture Search via Self-Evaluated Template Network. In Proceedings of the IEEE International Conference on Computer Vision, pages 3681–3690, 2019. 7

[12] Xuanyi Dong and Yi Yang. Searching for A Robust Neural Architecture in Four GPU Hours. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1761–1770, 2019. 7

[13] Xuanyi Dong and Yi Yang. NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search. In International Conference on Learning Representations, 2020. 1, 5, 7

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012. 6

[15] 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. European Conference on Computer Vision, 2020. 1, 2, 3, 4, 5, 6, 7, 11

[16] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. GhostNet: More Features from Cheap Operations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 6

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. Neural computation, 9(8):1735–1780, 1997. 2

[18] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for MobileNetV3. In International Conference on Computer Vision, 2019. 6, 7

[19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv preprint arXiv:1704.04861, 2017. 7

[20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7132–7141, 2018. 5, 6

[21] Yanping Huang, Yonglong Cheng, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, and Zhiheng Chen. GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism. Neural Information Processing Systems, 2019. 6

[22] Maurice G Kendall. A New Measure of Rank Correlation. Biometrika, 30(1/2):81–93, 1938. 3, 7

[23] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do Better ImageNet Models Transfer Better? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2661–2671, 2019. 6

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. Technical report, Citeseer, 2009. 2

[25] Liam Li and Ameet Talwalkar. Random Search and Reproducibility for Neural Architecture Search. Conference on Uncertainty in Artificial Intelligence, 2019. 2

[26] Hanwen Liang, Shifeng Zhang, Jiacheng Xu, Weiran Huang, Kechen Zhanga, and Zhenguo Li. DARTS+: Improved Differentiable Architecture Search with Early Stopping. arXiv preprint arXiv:1909.06035, 2019. 2

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In European Conference on Computer Vision, 2014. 6, 13

[28] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable Architecture Search. In International Conference on Learning Representations, 2019. 6

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In International Conference on Neural Information Processing Systems, 2012. 6

[30] Andrew Ng. Feature Maps for Deep Learning. In International Conference on Machine Learning, 2012. 6

[31] Yiping Liu, Ali R. Zamani, Zhijing Wang, and Yichen Wei. More Robust and Compact MobileNets with Iterative Optimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6

[32] Jiankang Lu, Xiangyu Zhang, Hao Li, Xiaogang Wang, and Yongfeng Yang. Rethinking ASSE: A New Network for the Robustness of Dense Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6

[33] Yixuan Li, Yichen Wei, and Yichen Wei. DAM: Differentiable Architecture Mining for Mobile Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 6

[34] Yu Cao, Yuxin Lin, Yi Zhang, and Jiaya Jia. AutoNet: Automating Design of Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6

[35] Han Cai, Ligeng Zhu, and Song Han. ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. In International Conference on Computer Vision, 2020. 6

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 6

[37] Andrew Ng. Feature Maps for Deep Learning. In International Conference on Machine Learning, 2012. 6

[38] Yiping Liu, Ali R. Zamani, Zhijing Wang, and Yichen Wei. More Robust and Compact MobileNets with Iterative Optimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6

[39] Jiankang Lu, Xiangyu Zhang, Hao Li, Xiaogang Wang, and Yongfeng Yang. Rethinking ASSE: A New Network for the Robustness of Dense Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6

[40] Yixuan Li, Yichen Wei, and Yichen Wei. DAM: Differentiable Architecture Mining for Mobile Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 6

[41] Han Cai, Ligeng Zhu, and Song Han. ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. In International Conference on Computer Vision, 2020. 6

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 6

[43] Andrew Ng. Feature Maps for Deep Learning. In International Conference on Machine Learning, 2012. 6

[44] Yiping Liu, Ali R. Zamani, Zhijing Wang, and Yichen Wei. More Robust and Compact MobileNets with Iterative Optimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6

[45] Jiankang Lu, Xiangyu Zhang, Hao Li, Xiaogang Wang, and Yongfeng Yang. Rethinking ASSE: A New Network for the Robustness of Dense Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6

[46] Yixuan Li, Yichen Wei, and Yichen Wei. DAM: Differentiable Architecture Mining for Mobile Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 6
A. More Discussion of Algorithms

A.1. Supernet Training

Figure 4 details the supernet training stage of our approach. In fact, it’s inherently efficient regarding GPU utilization. Even on powerful machines such as Tesla V100, it can make full use of GPU without special optimization. As most of the existing deep learning frameworks allow paralleled execution between data generation and gradient calculation, our algorithm can exploit this parallelism to the extreme since a mini-batch of data is reused by \( m \) times of backpropagation. The GPUs are always busy because the data is ready whenever required, which shortens the training time. Moreover, our method works in a single-path way, which is memory friendly.

Irregular Search Spaces. Note that SF in the paper can be easily extended by a preprocessing function in case of irregular search spaces (i.e. the number of operations are not the same for each layer). We only need to make a minor modification of Algorithm 1. Say the \( l \)-th layer has \( m_l \) choices. Suppose \( M = \max(m_l) \), we randomly choose \( M - m_l \) extra operations from \( m_l \) choices and regard these extra options as different ones from the original search space. Therefore, the input condition of Algorithm 1 still hold and we can use it directly. This procedure can be regarded as an approximated SF. However, perfect SF for irregular cases remains as our future work.

A.2. Evolutionary Searching Pipeline

With our supernet fairly trained as a model evaluator, we adopt an evolutionary-based algorithm for searching, detailed in Algorithm 2 (main text) and Figure 7. Generally, it is built on the ground of MoreMNAS [7] by replacing its incomplete-training evaluator with our fairly trained supernet. FairNAS supernet exhibits tremendous speed-up in terms of GPU days by two orders of magnitudes. We also use Proximal Policy Optimization as the default reinforcement algorithm [40].

![Figure 7. Evolutionary searching with the supernet trained with strict fairness. In each generation, candidate models in the current population inherit weights from the supernet for evaluation. Their estimated accuracies are fed into the searching pipeline as one of the objectives. The evolution loops till Pareto optimality.](image)

| NAS Methods | Type       | \( C_1 \) | \( C_s \) | EF | SF |
|------------|------------|--------|--------|----|----|
| SMASH [2]  | Hypernet   | -      | -      | \( \times \) | \( \times \) |
| One-Shot [1] | Supernet   | 4\( ^\dagger \) | 3.3   | \( \times \) | \( \times \) |
| DARTS [29] | Gradient-based | 0.5\( ^\ddagger \) | 0 | \( \times \) | \( \times \) |
| FBNet [49] | Gradient-based | 9 | 0 | \( \times \) | \( \times \) |
| ProxylessNAS [4] | Gradient-based/RL | 8.3 | 0 | \( \times \) | \( \times \) |
| SPOS [15] | Supernet+EA | 12 | <1 | \( \times \) | \( \times \) |
| Single-Path NAS [42] | Gradient-based | 1.25\( ^\ddagger \) | 0 | \( \checkmark \) | \( \times \) |
| FairNAS (Ours) | Fair Supernet+EA | 10 | 2 | \( \checkmark \) | \( \checkmark \) |

Table 7. Comparison of state-of-the-art weight-sharing NAS methods as per cost and fairness basis. \( C_1, C_s \): train and search cost measured in GPU days. EF: Expectation Fairness, SF: Strict Fairness. \( ^\ddagger \): searched on CIFAR-10, \( ^\dagger \): TPU

B. A Fairness Taxonomy

We compare current weight-sharing NAS methods based on the defined fairness in Table 7. SPOS [15] satisfies Expectation Fairness, while FairNAS meets Strict Fairness.

C. Experiment Details

Dataset. The supernet experiments are performed on ImageNet [38] and we randomly select 50,000 images from the training set as our validation set (50 samples from each class). The remaining training set is used as our training set, while the original validation set is taken as the test set to measure the final performance of each model.

C.1. Architectures of Searched Models

The searched FairNAS-A, B and C models are illustrated in Figure 8.

C.2. Hyperparameters for MoreMNAS variant

We list the hyperparameters for the adopted MoreMNAS [7] variant in Table 8. It has a population \( N \) of 64 models. It also takes a hierarchical mutation strategy. Respectively, \( p_{rm}, p_{re}, p_{pr} \) indicate probabilities for random mutation, reinforce mutation and prior regulator, where \( p_{re} \) again is divided into \( p_{K-M} \) for roulette wheel selection, and \( p_M \) for reinforced controller.

| Item                  | value |
|-----------------------|-------|
| Population N          | 64    |
| Mutation Ratio        | 0.8   |
| \( p_{rm} \)          | 0.2   |
| \( p_{re} \)          | 0.65  |
| \( p_{pr} \)          | 0.15  |
| \( p_M \)             | 0.7   |
| \( p_{K-M} \)         | 0.3   |

Table 8. Hyperparameters for the second-stage EA search.

C.3. Training of stand-alone models

We picked 13 models to train from scratch whose one-shot accuracies are approximately evenly spaced, ranging in
We keep the exactly same hyperparameters as the supernet training. Their corresponding stand-alone accuracies are within $[0.692, 0.715]$. Figure 9 plots the training process, from which we observe the ranking of one-shot models are generally maintained. The model-meta (indices of operations) of these 13 models are listed in Table 9. Besides, the mapping from an index in model-meta to a searchable operation is given in Table 10.

### Table 9. Model-meta of 13 sampled stand-alone models for ranking

| Model Meta Index | Model Meta     |
|------------------|----------------|
| 0                | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| 1                | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0] |
| 2                | [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0] |
| 3                | [0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0] |
| 4                | [0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0] |
| 5                | [0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0] |
| 6                | [0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 2, 0, 0] |
| 7                | [0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 2, 0, 0] |
| 8                | [0, 1, 4, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 2, 0, 0] |
| 9                | [0, 1, 0, 0, 1, 5, 1, 1, 0, 5, 1, 1, 1, 1, 1, 1] |
| 10               | [3, 1, 4, 1, 3, 4, 1, 4, 1, 3, 1, 3, 1, 2, 0, 0] |
| 11               | [0, 1, 4, 3, 1, 3, 1, 1, 1, 3, 4, 1, 3, 1, 2, 3] |
| 12               | [1, 5, 3, 2, 1, 4, 3, 4, 1, 5, 1, 1, 3, 5, 5, 3] |

### Table 10. Mapping between model-meta index and operations

| Model Meta Index | Kernel | Expansion Rate |
|------------------|--------|----------------|
| 0                | 3      | 3              |
| 1                | 5      | 6              |
| 2                | 7      | 3              |
| 3                | 3      | 6              |
| 4                | 5      | 6              |
| 5                | 7      | 6              |

Figure 9. Train and validation accuracies (ground truth) of all 13 stand-alone models when being fully trained with the same hyperparameters. Lines are labelled with corresponding one-shot accuracies (predicted) sorted in descending order (as reflected by color gradient).

### 4. Evolutionary Searching

The evolutionary search of FairNAS based on MoreMNAS variant [7] is shown in Figure 10. At each generation, 64 models are evaluated by our fair supernet, after 200 generations, the evolution converges, the Pareto-front is shown in bright yellow, each dot represents a candidate network.
C.5. Object Detection

For object detection, we treat FairNAS models as drop-in replacements for RetinaNet's backbone [27]. We follow the same setting as [27] and exploit MMDetection toolbox [5] for training. All the models are trained and evaluated on MS COCO dataset (train2017 and val2017 respectively) [28] for 12 epochs with a batch size of 16. The initial learning rate is 0.01 and decayed by 0.1 × at epochs 8 and 11.

All baselines in the paper are mobile networks. The input features from these backbones to the FPN module are from the last depthwise layers of stage 2 to 5 5. The number of output channels of FPN is kept 256 as [27]. We also use $\alpha = 0.25$ and $\gamma = 2.0$ for the focal loss. Given longer training epochs and other tricks, the detection performance can be improved further. However, it's sufficient to compare the transferability of various methods.

---

5We follow the typical nomination for the definition of stages and the orders start from 1.