Tackling Neural Architecture Search With Quality Diversity Optimization

Anonymous[1]

[1]Anonymous Institution

Abstract Neural architecture search (NAS) has been studied extensively and has grown to become a research field with substantial impact. While classical single-objective NAS searches for the architecture with the best performance, multi-objective NAS considers multiple objectives that should be optimized simultaneously, e.g., minimizing resource usage along the validation error. Although considerable progress has been made in the field of multi-objective NAS, we argue that there is some discrepancy between the actual optimization problem of practical interest and the optimization problem that multi-objective NAS tries to solve. We resolve this discrepancy by formulating the multi-objective NAS problem as a quality diversity optimization (QDO) problem and introduce three quality diversity NAS optimizers (two of them belonging to the group of multifidelity optimizers), which search for high-performing yet diverse architectures that are optimal for application-specific niches, e.g., hardware constraints. We compare these optimizers to their multi-objective counterparts and demonstrate that quality diversity NAS in general outperforms multi-objective NAS with respect to quality of solutions and efficiency. Using model compression, we show how QDO can be applied to obtain a set of sparser yet high-performing architectures.

1 Introduction

The goal of neural architecture search (NAS) is to automate the manual process of designing optimal neural network architectures. Traditionally, NAS is formulated as a single-objective optimization problem with the goal of finding an architecture that has minimal validation error [10, 31, 41, 43, 42, 56]. Considerations for additional objectives such as efficiency have led to the formulation of constraint NAS methods that enforce efficiency thresholds [1] as well as multi-objective NAS methods [7, 9, 33, 47, 32] that yield a Pareto optimal set of architectures. However, in most practical applications, we are not interested in the complete Pareto optimal set. Instead, we would like to obtain solutions for a discrete set of scenarios (e.g., end-user devices), which we henceforth refer to as *niches* in this paper. This is illustrated in Figure 1.

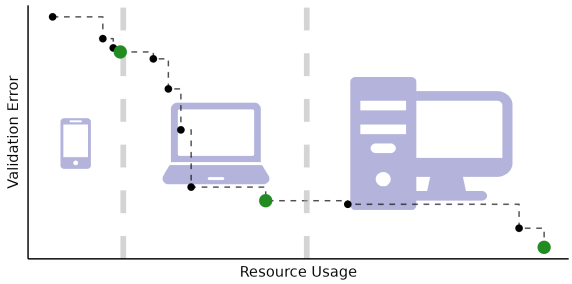

Figure 1: Optimizing neural network architectures for a discrete set of devices simultaneously. We are interested in the best solution (green) within the constraints of the respective device (vertical dashed lines). Multi-objective optimization searches the full Pareto front (black).

 © 2022 the authors, released under CC BY 4.0

A concrete example is finding neural architectures for microcontrollers [28] and other edge devices [34], e.g., in μNAS [28] architectures for "mid-tier" IoT devices with up to 64 KB of SRAM and 64 KB of storage are searched. To evaluate the benefits for larger devices, the search would need to be restarted with adapted constraints, thus wasting computational resources. Formulating the search as a multi-objective problem would also waste resources; once an architecture satisfies the constraints of a device, we are not interested in additional trade-offs, and we select only based on the validation error. If the goal is to find architectures that can use up to a certain amount of memory so they can later be deployed on different devices, it is not beneficial to further reduce the resource objective at the cost of worse performance once the constraint of the device is met. A similar argument can be made for inference latency.

We therefore argue that the multi-objective NAS problem *can* and usually *should* be formulated as a *quality diversity optimization* (QDO) problem, which directly corresponds to the actual optimization problem of interest. The main contributions of this paper are: We (1) formulate multi-objective NAS as a QDO problem; (2) show how to adapt black-box optimization algorithms for the QDO setting; (3) modify existing QDO algorithms for the NAS setting; and (4) propose novel multifidelity QDO algorithms for NAS. Our benchmark results demonstrate that QDO outperforms multi-objective optimization in terms of quality and efficiency.

2 Theoretical Background and Related Work

Let $\mathcal{A}$ denote the search space of architectures and Λ the search space of additional hyperparameters. Furthermore, let $f_{\text{err}} : \mathcal{A} \times \Lambda \to \mathbb{R}$ denote the validation error obtained after training an architecture $A \in \mathcal{A}$ with a set of hyperparameters $\lambda \in \Lambda$ for a given number of epochs ($\lambda_{\text{epoch}} \in \lambda$). Typically, we consider $\lambda \in \Lambda$ to be fixed, except for λ_{epoch} in multi-fidelity methods, and omit λ in the following. The goal of single-objective NAS is to find the architecture with the lowest validation error, $A^* := \arg\min_{A \in \mathcal{A}} f_{\text{err}}(A)$.

NAS methods can be categorized along three dimensions: search space, search strategy, and performance estimation strategy [10]. For chain-structured neural networks (simply a connected sequence of layers), cell-based search spaces have gained popularity [42, 41]. In cell-based search spaces, different kinds of cells – typically, a normal cell preserving dimensionality of the input and a reduction cell reducing spatial dimension – are stacked in a predefined arrangement to form a final architecture. Regarding search strategy, popular methods utilize Bayesian Optimization (BO) [2, 5, 35, 20, 52], evolutionary methods [37, 30, 43, 42, 9], reinforcement learning [56, 57], or gradient-based algorithms [31, 41]. For performance estimation, popular approaches leverage lower fidelity estimates [27, 11, 57] or make use of learning curve extrapolation [5, 23].

Multi-Objective Neural Architecture Search Contrary to the single-objective NAS formulation, multi-objective NAS does not solely aim for minimizing the validation error but simultaneously optimizes multiple objectives. These objectives typically take resource consumption – such as memory requirements, energy usage or latency – into account [7, 9, 33, 47, 32]. Denote by $f_1, \ldots, f_k$ the $k \geq 2$ objectives of interest, where typically $f_1 = f_{\text{err}}$ and denote by $\mathbf{f}(A)$ the vector of objective function values obtained for architecture $A \in \mathcal{A}$, $\mathbf{f}(A) := (f_1(A), \ldots, f_k(A))'$. The optimization problem of multi-objective NAS is then formulated as $\min_{A \in \mathcal{A}} \mathbf{f}(A)$. A solution to the problem is given by a set of so-called Pareto optimal architectures that reflect different trade-offs in objectives approximating the true (unknown) Pareto front. An architecture A is said to dominate another architecture A' iff $\forall i \in \{1, \ldots, k\} : f_i(A) \leq f_i(A') \wedge \exists j \in \{1, \ldots, k\} : f_j(A) < f_j(A')$.

Constrained and Hardware-Aware Neural Architecture Search In contrast, *Constrained NAS* [55, 12, 49] solves the problem of finding an architecture that optimizes one objective (e.g., validation error) with constraints on secondary objectives (e.g., model size). Constraints can be naturally given by the target hardware that a model should be deployed on. *Hardware-Aware NAS* in turn searches for an architecture that trades off primary objectives [53] against secondary, hardware-specific metrics. For a recent survey, we refer to [1].

Quality Diversity Optimization The goal of a QDO algorithm is to find a set of high-performing, yet behaviorally diverse, solutions. Similarly to multi-objective optimization, there is no single best solution. However, whereas multi-objective optimization aims for the simultaneous minimization of multiple objectives, QDO minimizes a single-objective function with respect to diversity defined on one or more *feature functions*. A feature function measures a quality of interest and a combination of feature values points to a niche, i.e., a region in *feature space*. QDO could be considered a *set* of constrained optimisation problems over the same input domain where the niche boundaries are constraints in feature space. The key difference is that constrained optimisation seeks a single optimal configuration given some constraints, while QDO attempts to identify the optimal configuration for each of a set of constrained regions simultaneously. In this sense, QDO could be framed as a so-called multi-task optimization problem [39] where each task is to find the best solution belonging to a particular niche.

QDO algorithms maintain an archive of niche-optimal observations, i.e., a best-performing observed solution for each niche. Observations with similar feature values compete to be selected for the archive, and the solution set gradually improves during the optimization process. Once the optimization budget has been spent, QDO algorithms typically return this archive as their solution. QDO is motivated by applications where a group of diverse solutions is beneficial, such as the training of robot movement where a repertoire of behaviours must be learned [4], developing game playing agents with diverse strategies [40], and in automatic design where QDO can be used by human designers to search a large dimensional search space for diverse solutions before the optimization is finished by hand. Work on automatic design tasks have been varied and include air-foil design [15], computer game level design [13], and architectural design [6].

In the earliest examples, Novelty Search (NS; [25]) asks whether diversity alone can produce a good set of solutions. Despite not actively pursuing objective performance, NS performed surprisingly well in some settings and was followed by Novelty Search with Local Competition [26], the first true quality diversity algorithm. MAP-Elites [38], a standard evolutionary QDO algorithm, partitions the feature space a-priori into niches and attempts to identify the optimal solution in each of these niches. QDO has seen much work in recent years and a variant based on BO, BOP-Elites, was proposed recently [21]. BOP-Elites models the objective and feature functions with surrogate models and implements an acquisition function over a structured archive to achieve high sample efficiency even in the case of black-box features.

3 Formulating Neural Architecture Search as a Quality Diversity Optimization Problem

In the example in Figure 1, a quality diversity NAS (subsequently abbreviated as qdNAS) problem is given by the validation error and three behavioral niches (corresponding to different devices) that are defined via resource usage measured by a single feature function. Let $f_1 : \mathcal{A} \to \mathbb{R}, A \mapsto f_1(A)$ denote the objective function of interest (in our context, f_{err}). Denote by $f_i : \mathcal{A} \to \mathbb{R}, A \mapsto f_i(A), i \in \{2, \dots, k\}, k \geq 2$ the feature function(s) of interest (e.g., memory usage). Behavioral niches $N_j \subseteq \mathcal{A}, j \in \{1, \dots, c\}, c \geq 1$ are sets of architectures characterized via niche-specific boundaries $\mathbf{b}_{ij} = \left[l_{ij}, u_{ij}\right) \subseteq \mathbb{R}$ on the images of the feature functions. An architecture A belongs to niche N_j if its values with respect to the feature functions lie between the respective boundaries, i.e.:

$$A \in N_j \iff \forall i \in \{2, \dots, k\} : f_i(A) \in \mathbf{b}_{ij}.$$

The goal of a QDO algorithm is then to find for each behavioral niche N_j the architecture that minimizes the objective function f_1:

$$\mathbf{A}_j^* := \underset{A \in N_j}{\arg\min}\, f_1(A).$$

In other words, the goal is to obtain a set of architectures $\mathcal{S} := \{A_1^*, \ldots, A_c^*\}$ that are diverse with respect to the feature functions and yet high-performing.

A Remark about Niches In the classical QDO literature, niches are typically constructed to be pairwise disjoint, i.e., a configuration can only belong to a single niche (or none) [38, 21]. However, depending on the concrete application, relaxing this constraint and allowing for *overlap* can be beneficial. For example, in our context, an architecture that fits on a mid-tier device should also be considered for deployment on a higher-tier device, i.e., in Figure 1, the boundaries indicated by vertical dashed lines resemble the respective upper bound of a niche whereas the lower bound is unconstrained. This results in niches being nested within each other, i.e., $N_1 \subsetneq N_2 \subsetneq \ldots \subsetneq N_c \subseteq \mathcal{A}$, with N_1 being the most restrictive niche, followed by N_2.

3.1 Quality Diversity Optimizers for Neural Architecture Search

As the majority of NAS optimizers are iterative, we first demonstrate how any iterative optimizer can in principle be turned into a quality diversity optimizer. Based on this correspondence, we introduce three novel quality diversity optimizers for NAS: BOP-Elites*, qdHB and BOP-ElitesHB. Let $f_1 : \mathcal{A} \to \mathbb{R}, A \mapsto f_1(x)$ denote the objective function that should be minimized. In each iteration, an iterative optimizer proposes a new configuration (e.g., architecture) for evaluation (e.g., based on a surrogate model), evaluates this configuration, potentially updates the incumbent (best configuration evaluated so far) if better performance has been observed, and updates its archive. For generic pseudo code, see **??**.

Moving to a QDO problem, there are now feature functions $f_i : \mathcal{A} \to \mathbb{R}, A \mapsto f_i(A)$, $i \in \{2, \ldots, k\}, k \geq 2$, and niches N_j, $j \in \{1, \ldots, c\}, c \geq 1$, defined via their niche-specific boundaries $\mathbf{b}_{ij} = \left[l_{ij}, u_{ij}\right) \subseteq \mathbb{R}$ on the images of the feature functions. Any iterative single-objective optimizer must then keep track of the best incumbent per niche (often referred to as the *elite* in the QDO literature) and essentially becomes a quality diversity optimizer (see Algorithm 1). The

Algorithm 1: Generic pseudo code for an iterative quality diversity optimizer.

Input : $f_1, f_i, i \in \{2, \ldots, k\}, k \geq 2, N_j, j \in \{1, \ldots, c\}, c \geq 1, \mathcal{D}_{\text{design}}, n_{\text{total}}$
Result: $S = \{A_1^*, \ldots, A_c^*\}$

```
D ← D_design
for j ← 1 to c do
    A*_j ← arg min_{A∈D|N_j} f_1(A) # initial incumbent of niche N_j based on archive
end
for n ← 1 to n_total do
    Propose a new candidate A* # subroutine
    Evaluate y ← f_1(A*), ∀i ∈ {2, ..., k} : z_i ← f_i(A*)
    if A* ∈ N_j ∧ y < f_1(A*_j) then
        A*_j ← A* # update incumbent of niche N_j
    end
    D ← D ∪ {(A*, y, z_2, ..., z_k)}
end
```

challenge in designing an efficient and well-performing quality diversity optimizer now mostly lies in proposing a new candidate for evaluation that considers improvement over all niches.

Bayesian Optimization A recently proposed model-based quality diversity optimizer – BOP-Elites [21] – extends BO [48, 14] to QDO. BOP-Elites relies on Gaussian process surrogate models for the objective function and all feature functions. New candidates for evaluation are selected by a novel acquisition function – the expected joint improvement of elites (EJIE), which measures the

expected improvement to the ensemble problem of identifying the best solution in every niche:

$$\alpha_{\text{EJIE}}(A) := \sum_{j=1}^{c} \mathcal{P}(A \in N_j | \mathcal{D}) \mathbb{E}_y \left[\mathbb{I}_{|N_j}(A) \right]. \tag{1}$$

Here, $\mathcal{P}(A \in N_j|\mathcal{D})$ is the posterior probability of A belonging to niche N_j, and $\mathbb{E}_y \left[\mathbb{I}_{|N_j}(A) \right]$ is the expected improvement (EI; [19]) with respect to niche N_j:

$$\mathbb{E}_y \left[\mathbb{I}_{|N_j}(A) \right] = \mathbb{E}_y \left[\max \left(f_{\min_{\mathrm{N_j}}} - y, 0 \right) \right],$$

where $f_{\min_{N_j}}$ is the best observed objective function value in niche N_j so far, and y is the surrogate model prediction for A. A new candidate is then proposed by maximizing the EJIE, i.e., Line 6 in Algorithm 1 looks like the following: $A^\star \leftarrow \arg\max_{A \in \mathcal{A}} \alpha_{\text{EJIE}}(A)$.

BOP-Elites* In order to adapt BOP-Elites for NAS, we introduce several modifications. First, we employ truncated path encoding [51, 52], transforming architectures into a set of binary features indicating the presence of all directed edges along paths. Second, we substitute the Gaussian process surrogate models used in BOP-Elites with random forests allowing us to model non-continuous hyperparameter spaces. Random forests have been successfully used as surrogates in BO [17, 29], often performing on a par with ensembles of neural networks [52] in the context of NAS [46, 50]. Third, we introduce a local mutation scheme similarly to the one used by the BANANAS algorithm [52] for optimizing the infill criterion EJIE: Since our aim is to find high quality solutions across all niches, we maintain an archive of the incumbent architecture in each niche and perform local mutations on each incumbent. We refer to our adjusted version as BOP-Elites* in the remainder of the paper to emphasize the difference from the original algorithm. For the initial design, we sample architectures based on adjacency matrix encoding [51].

Multi-fidelity Optimizers For NAS, performance estimation is the computationally most expensive component [10], and almost all NAS optimizers can be made more efficient by allowing access to cheaper, lower fidelity estimates [10, 27, 11, 57]. By evaluating most architectures at lower fidelity and only promoting promising architectures to higher fidelity, many more architectures can be explored given the same total computational budget. The fidelity parameter is typically the number of epochs over which an architecture is trained.

qdHB One of the most prominent multifidelity optimizers is Hyperband (HB; [27]), a multi-armed bandit strategy that uses repeated Successive Halving (SH) [18] as a subroutine to identify the best configuration (e.g., architecture) among a set of randomly sampled ones. Given an initial and maximum fidelity, a scaling parameter η, and a set of configurations of size n, SH evaluates all configurations on the initial smallest fidelity, then sorts the configurations by performance and only keeps the best $1/\eta$ configurations. These configurations are then trained with fidelity increased by a factor of η. This process is repeated until the maximum fidelity for a single configuration is reached. HB repeatedly runs SH with different sized sets of initial configurations called brackets. Only two inputs are required: R, the maximum fidelity and η, the scaling parameter that controls the proportion of configurations discarded in each round of SH. Based on these inputs, the number $s_{\max}$ and size n_i of different brackets is determined. To adapt HB to the quality diversity setting, we must track the incumbent architecture in each niche and promote configurations based on their performance within the respective niche (see **??**): To achieve this, we choose the top $\lfloor n_i/\eta \rfloor$ configurations to be promoted uniformly over the c niches (done in the $\texttt{top}_{\texttt{k_qdo}}$ function), i.e., we iteratively select one of the niches uniformly at random and choose the best configuration observed so far that has yet not been selected for promotion until $\lfloor n_i/\eta \rfloor$ configurations have been selected. Note that during this procedure, it may happen that not enough configurations belonging to a specific niche have been observed yet. In this case, we choose any configuration uniformly at

random over the set of all configurations that have yet to be promoted. With those modifications, we propose qdHB, as a multifidelity quality diversity optimizer.

BOP-ElitesHB While HB typically shows strong anytime performance [27], it only samples configurations at random and is typically outperformed by BO methods with respect to final performance if optimizer runtime is sufficiently large [11]. BOHB [11] combines the strengths of HB and BO in a single optimizer, resulting in strong anytime performance and fast convergence. This approach employs a fidelity schedule similar to HB to determine how many configurations to evaluate at which fidelity but replaces the random selection of configurations in each HB iteration by a model-based proposal. In BOHB, a Tree Parzen Estimator [3] is used to model densities $l(A) = p(y < \alpha | A, \mathcal{D})$ and $g(A) = p(y > \alpha | A, \mathcal{D})$, and candidates are proposed that maximize the ratio $l(A)/g(A)$, which is equivalent to maximizing EI [3]. In contrast, other multifidelity optimizers combining HB and BO differ in their choice of surrogate model and implementation detail, e.g., SMAC-HB [29] uses a random forest. Based on BOP-Elites* and qdHB, we can now derive the quality diversity Bayesian Optimization Hyperband optimizer (BOP-ElitesHB): Instead of selecting configurations at random at the beginning of each qdHB iteration, we propose candidates that maximize the EJIE criterion. This sampling procedure is described in **??**.

4 Benchmark Experiments and Results

We are interested in answering the following research questions: (**RQ1**) *Does qdNAS outperform multi-objective NAS if the optimization goal is to find high-performing architectures in pre-defined niches?* (**RQ2**) *Do multifidelity qdNAS optimizers improve over full-fidelity qdNAS optimizers?* To answer these questions, we benchmark our three qdNAS optimizers – BOP-Elites*, qdHB, and BOP-ElitesHB– on the well-known NAS-Bench-101 [54] and NAS-Bench-201 [8] and compare them to three multi-objective optimizers adapted for NAS: ParEGO*, moHB*, and ParEGOHB as well as a simple Random Search[1].

Experimental Setup It is important to compare optimizers using analogous implementation details. We therefore use truncated path encoding and random forest surrogates throughout our experiments for all model-based optimizers. Furthermore, we use local mutations as described in [52] in order to optimize acquisition functions in BOP-Elites*, BOP-ElitesHB, ParEGO*, and ParEGOHB. To control for differences in implementation, we re-implement all optimizers and take great care in matching the original implementations.

We provide full details regarding implementation in **??** and only briefly introduce conceptual differences: ParEGO* is a multi-objective optimizer based on ParEGO [24] and only deviates from BOP-Elites* in that it considers a differently scalarized objective in each iteration, which is optimized using local mutations similar to the acquisition function optimization of BOP-Elites*. moHB* is an extension of HB to the multi-objective setting (promoting configurations based on non-dominated sorting with hypervolume contribution for tie breaking, see, e.g., [44, 45, 16]). ParEGOHB is a model-based extension of moHB* (see, e.g., [44]) that relies on the ParEGO scalarization [24] and on the same acquisition function optimization as ParEGO*.

All optimizers were evaluated on NAS-Bench-101 (Cifar-10, with the validation error as the first objective and the number of trainable parameters as the feature function / second objective) and NAS-Bench-201 (Cifar-10, Cifar-100, ImageNet16-120, with the validation error as the first objective and latency as the feature function / second objective). For multifidelity, we train architectures for 4, 12, 36, 108 epochs on NAS-Bench-101 and for 2, 7, 22, 67, 200 epochs on NAS-Bench-201 (reflecting $\eta = 3$ in the HB variants). As the optimization budget, we consider 200 full architecture evaluations (resulting in a total budget of 21600 epochs for NAS-Bench-101 and 40000 epochs for NAS-Bench-201). For each of these four settings, we construct three different scenarios by considering different niches of interest with respect to the feature function, resulting in a total of 12 benchmark prob-

[1] using adjacency matrix encoding [51]

lems. In the *small/medium/large* settings, two, five and ten niches are considered, respectively. Niches are constructed to be overlapping, and boundaries are defined based on percentiles of the feature function. For the *small* setting, the boundary is given by the 50% percentile, effectively resulting in two niches with boundaries $[0, q_{50})$ and $[0, \infty)$. For the *medium* and *large* settings, the following percentiles indicating progressively larger niches were used: 1%, 5%, 10%, 30% and 1%, 2%, 5%, 10%, 20%, 30%, 40%, 50%, 70%, respectively. More details on the niches can be found in **??**. All runs are replicated 100 times.

Results As an anytime performance measure, we are interested in the validation error obtained for each niche, which we aggregate in a single performance measure as $\sum_{j=1}^{c} f_{\text{err}}(A_j^*)$, i.e., we consider the sum of validation errors over the best-performing architecture per niche. If a niche is empty, we assign a validation error of 100 as a penalty. For the final performance, we also consider the analogous test error. Figure 2 shows the anytime performance of optimizers with respect to the validation error summed over niches (averaged over 100 replications). We observe that model-based optimizers (BOP-Elites* and ParEGO*) in general strongly outperform Random Search, and BO HB optimizers (BOP-ElitesHB and ParEGOHB) generally outperform their full-fidelity counterparts, although this effect diminishes with increasing optimization budget. In general, HB variants that do not rely on a surrogate model (qdHB and moHB*) show poor performance compared to the model-based optimizers. Moreover, especially in the *small* number of niches setting, quality diversity strongly outperforms multi-objective optimization. Mean ranks of optimizers with respect to final validation and test performance are given in Table 1. For completeness, we also report critical differences plots of these ranks in **??**. Additionally, we report the average best validation and test performance for each niche for each optimizer on each benchmark problem in **??**.

We also conducted two four-way ANOVAS on the average performance after half and all of the optimization budget is used, with the factors problem (benchmark problem), multifidelity (whether an optimizer uses multifidelity), QDO (whether an optimizer is a QDO optimizer) and model-based (whether the optimizer relies on a surrogate model)[2]. For half the budget used, results indicate significant main effects of the factors multifidelity ($F(1) = 19.13, p = 0.0001$), QDO ($F(1) = 11.08, p = 0.0017$) and model-based ($F(1) = 21.13, p < 0.0001$). For all of the budget used, the significance of multifidelity diminishes, whereas the main effects of QDO ($F(1) = 18.31, p = 0.0001$) and model-based ($F(43.44), p < 0.0001$) are still significant. We can therefore conclude that QDO in general results in significantly better performance when the goal is to find high-performing architectures in pre-defined niches. Detailed results are reported in **??**.

Regarding efficiency, we analyzed the expected running time (ERT) of the QDO optimizers given the average performance of the respective multi-objective optimizers after half of the optimization budget: For each benchmark problem, we computed the mean validation performance of each multi-objective optimizer after having spent half of its optimization budget and investigated the ERT of the analogous[3] QDO optimizer. We scaled the ERTs by the total optimization budget per benchmark problem and then computed the ratio of ERTs between multi-objective and QDO optimizers and averaged them over the benchmark problems. For BOP-ElitesHB, we observe an average ERT ratio of 2.41, i.e., in expectation, BOP-ElitesHB is a factor of 2.41 faster than ParEGOHB in reaching the average performance of ParEGOHB (after half the optimization budget). For qdHB and BOP-Elites*, the average ERT ratios are 1.14 and 1.44. We conclude that all QDO optimizers are more efficient than their multi-objective counterparts. More details on this analysis can be found in **??**.

5 Applying Quality Diversity Neural Architecture Search to Model Compression

We now demonstrate how qdNAS can be used in practice and consider the following scenario: We are interested in deploying a `MobileNetV2` across different devices that are mainly constrained

[2]For this analysis, we excluded qdHB and moHB* due to their lackluster performance.
[3]BOP-ElitesHB for ParEGOHB, qdHB for moHB*, and BOP-Elites* for ParEGO*

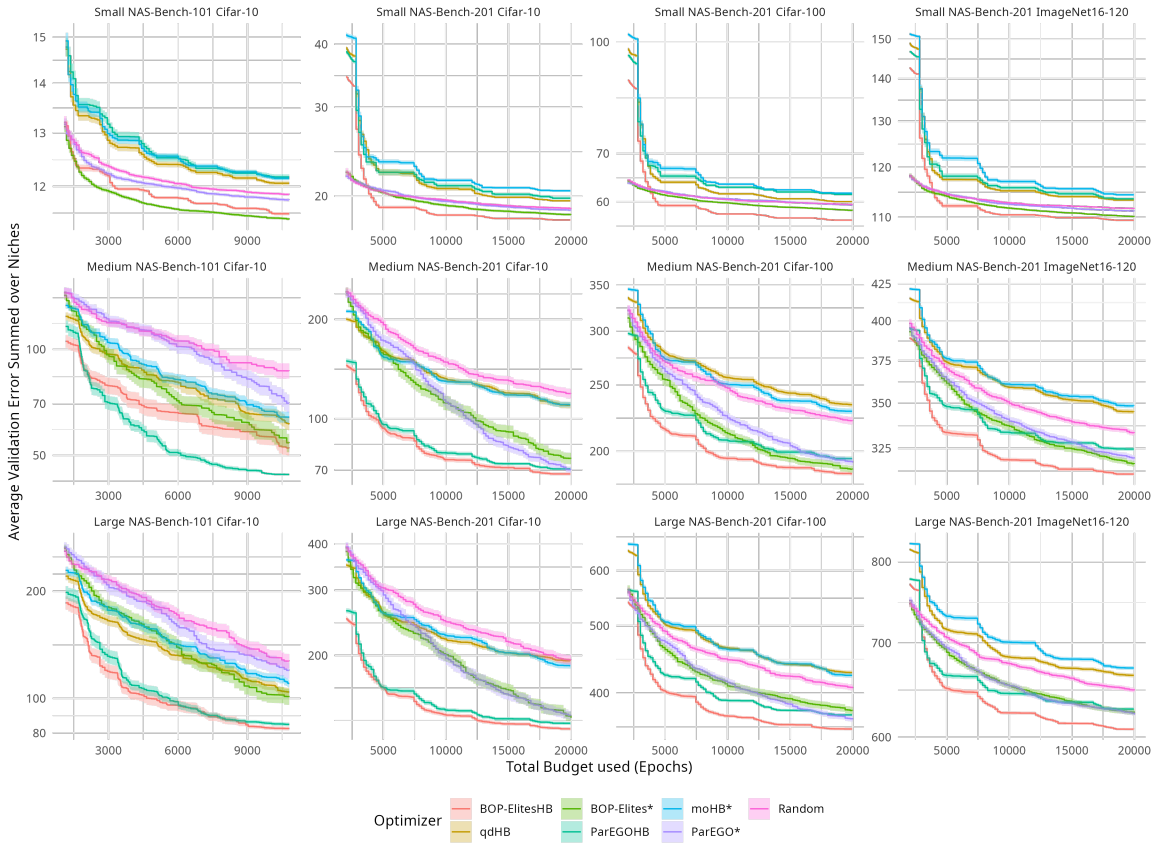

Figure 2: Anytime performance of optimizers (validation error summed over niches). Ribbons represent standard errors over 100 replications. x-axis starts after 10 full-fidelity evaluations.

Table 1: Mean ranks of optimizers with respect to their final validation and test performance, averaged over all 12 benchmark problems.

Mean Rank (SE)	BOP-ElitesHB	qdHB	BOP-Elites*	ParEGOHB	moHB*	ParEGO*	Random
Validation	2.08 (0.29)	5.92 (0.26)	1.83 (0.21)	4.25 (0.48)	6.42 (0.15)	2.58 (0.31)	4.92 (0.34)
Test	1.42 (0.19)	5.00 (0.17)	2.08 (0.23)	4.33 (0.50)	6.50 (0.19)	3.25 (0.35)	5.42 (0.42)

by memory. For each device, we can therefore only consider models up to a fixed amount of parameters, similar to the scenario depicted in Figure 1. If a model fits into the memory of a given device, we are not interested in further available trade-offs between memory and performance. Given that we have a pretrained model that achieves high performance, we want to *compress* this model exploiting redundancies in model parameters. In our application, we use the Stanford Dogs dataset [22] and rely on the neural network intelligence (NNI; [36]) toolkit for model compression. Pruning consists of several (iterative) steps as well as re-training of the pruned architectures. Several choices for the pruner itself, pruner hyperparameters, and hyperparameters controlling retraining are available and must be carefully selected in order to obtain optimal models. We consider the number of model parameters as a proxy measure for memory requirement, yielding three overlapping niches for different devices. The pre-trained `MobileNetV2` achieves a validation error of 20.25% using 2343480 model parameters. We define niches with the following boundaries: $[937392, 1171740)$, $[937392, 1406088)$ and $[937392, 1640436)$ corresponding to compression rates (number of parameters after pruning) of 40% to 50%, 40% to 60%, and 40% to 70%. The complete

search space is given in **??**. As the quality diversity optimizer, we decide to use BOP-ElitesHB and specify the number of fine-tuning epochs a pruner can use as a fidelity parameter, since fine-tuning after pruning is costly but also strongly influences final performance. After evaluating only 69 pruning configurations (a single BOP-ElitesHB run with $\eta = 3$), we obtain high-performing pruner configurations for each niche, resulting in the performance vs. memory requirement (number of parameters) trade-off shown in Table 2.

Table 2: Results of using BOP-ElitesHB for model compression of `MobileNetV2` on Stanford Dogs.

Niche	**Validation Error**	**# Param**	**% (# Param$_{\text{Baseline}}$)**
Niche 1 $[937392, 1171740)$	31.20%	1127166	48.10%
Niche 2 $[937392, 1406088)$	29.07%	1288582	54.99%
Niche 3 $[937392, 1640436)$	27.76%	1616339	68.97%
Baseline	20.25%	2343480	100%

6 Conclusion

We demonstrated how multi-objective NAS can be formulated as a QDO problem that, contrary to multi-objective NAS, directly corresponds to the actual optimization problem of interest, i.e., finding high-performing architectures in different niches. We have shown how any iterative black box optimization algorithm can be adapted to the QDO setting and proposed three QDO algorithms for NAS, with two of which making use of multifidelity evaluations. In benchmark experiments, we have shown that qdNAS outperforms multi-objective NAS while simultaneously being more efficient. We furthermore illustrated how qdNAS can be used for model compression.

Limitations The framework we describe relies on pre-defined niches, e.g., memory requirements of different devices. If niches are mis-specified or cannot be specified a priori, multi-objective NAS may outperform qdNAS. However, an initial study (see **??**) how qdNAS performs in a true multi-objective setting, which would correspond to unknown niches, shows little to no performance degradation depending on the choice of niches. Moreover, we only investigated the performance of qdNAS in the deterministic setting. Additionally, our multifidelity optimizers require niche membership to be unaffected by the multifidelity parameter (which is typically not problematic in the case of resource-related feature functions). Finally, we mainly focused on model-based NAS algorithms that we have extended to the QDO setting. Future work may aim to address the question of how to automatically specify niches if a priori information is not available, extend qdNAS to the noisy setting and adapt other NAS algorithms relying on different search strategies to the QDO setting.

Broader Impact Our work extends previous research on NAS and therefore inherits its implications on society and individuals such as potential discrimination in resulting models. Moreover, evaluating a large number of architectures is computationally costly and can introduce serious environmental issues. We have shown that qdNAS allows for finding better solutions, while simultaneously being more efficient than multi-objective NAS. As performance estimation is extremely costly in NAS, we believe that this is an important contribution towards reducing resource usage and the CO_2 footprint of NAS.

7 Reproducibility Checklist

1. For all authors...

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

   (b) Did you describe the limitations of your work? [Yes] See Section 6.

   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.

   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [N/A]

   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results, including all requirements (e.g., `requirements.txt` with explicit version), an instructive `README` with installation, and execution commands (either in the supplemental material or as a URL)? [Yes] The full code for experiments, application, figures and tables can be obtained from the following GitHub repository: `https://anonymous.4open.science/r/qdo_nas-7383`.

   (b) Did you include the raw results of running the given instructions on the given code and data? [Yes] Raw results are provided via the same GitHub repository.

   (c) Did you include scripts and commands that can be used to generate the figures and tables in your paper based on the raw results of the code, data, and instructions given? [Yes] Scripts to generate figures and tables based on raw results are provided via the same GitHub repository.

   (d) Did you ensure sufficient code quality such that your code can be safely executed and the code is properly documented? [Yes]

   (e) Did you specify all the training details (e.g., data splits, pre-processing, search spaces, fixed hyperparameter settings, and how they were chosen)? [Yes] For our benchmark experiments we used NAS-Bench-101 and NAS-Bench-201. For the application of QDO to model compression all details are reported in **??**.

   (f) Did you ensure that you compared different methods (including your own) exactly on the same benchmarks, including the same datasets, search space, code for training and hyperparameters for that code? [Yes] As described in Section 4.

   (g) Did you run ablation studies to assess the impact of different components of your approach? [Yes] See **??**, **??** and **??**.

   (h) Did you use the same evaluation protocol for the methods being compared? [Yes] As described in Section 4.

   (i) Did you compare performance over time? [Yes] Anytime performance was assessed with respect to the number of epochs as described in Section 4.

(j) Did you perform multiple runs of your experiments and report random seeds? [Yes] All runs were replicated 100 times. Random seeds can be obtained via `https://anonymous.4open.science/r/qdo_nas-7383`.

(k) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All results include error bars accompanying mean estimates.

(l) Did you use tabular or surrogate benchmarks for in-depth evaluations? [Yes] We used the tabular NAS-Bench-101 and NAS-Bench-201 benchmarks.

(m) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] As described in **??**.

(n) Did you report how you tuned hyperparameters, and what time and resources this required (if they were not automatically tuned by your AutoML method, e.g. in a NAS approach; and also hyperparameters of your own method)? [N/A] No tuning of hyperparameters was performed.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets…

(a) If your work uses existing assets, did you cite the creators? [Yes] NAS-Bench-101, NAS-Bench-201, Naszilla, NNI and the Stanford Dogs dataset are cited appropriately.

(b) Did you mention the license of the assets? [Yes] Done in **??**.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We provide all our code via `https://anonymous.4open.science/r/qdo_nas-7383`.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All assets used are either released under the Apache-2.0 License or MIT License.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects…

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

References

[1] Hadjer Benmeziane, Kaoutar El Maghraoui, Hamza Ouarnoughi, Smail Niar, Martin Wistuba, and Naigang Wang. Hardware-aware neural architecture search: Survey and taxonomy. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4322–4329, 2021.

[2] J. Bergstra, D. Yamins, and D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning*, pages 115–123, 2013.

[3] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[4] A. Cully and J.-B. Mouret. Evolving a behavioral repertoire for a walking robot. *Evolutionary Computation*, 24(1):59–88, 2016.

[5] T. Domhan, J. T. Springenberg, and F. Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3460—-3468, 2015.

[6] Stephane Doncieux, Nicolas Bredeche, Léni Le Goff, Benoît Girard, Alexandre Coninx, Olivier Sigaud, Mehdi Khamassi, Natalia Díaz-Rodríguez, David Filliat, Timothy Hospedales, et al. Dream architecture: A developmental approach to open-ended learning in robotics. *arXiv preprint arXiv:2005.06223*, 2020.

[7] Jin-Dong Dong, An-Chieh Cheng, Da-Cheng Juan, Wei Wei, and Min Sun. Dpp-net: Device-aware progressive search for pareto-optimal neural architectures. In *European Conference on Computer Vision*, 2018.

[8] X. Dong and Y Yang. NAS-Bench-201: Extending the scope of reproducible neural architecture search. In *International Conference on Learning Representations*, 2020.

[9] T. Elsken, J. H. Metzen, and F. Hutter. Efficient multi-objective neural architecture search via Lamarckian evolution. In *Proceedings of the International Conference on Learning Representations*, 2019.

[10] T. Elsken, J. H. Metzen, and F. Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21, 2019.

[11] Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[12] Igor Fedorov, Ryan P Adams, Matthew Mattina, and Paul Whatmough. Sparse: Sparse architecture search for CNNs on resource-constrained microcontrollers. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Matthew C Fontaine, Ruilin Liu, Julian Togelius, Amy K Hoover, and Stefanos Nikolaidis. Illuminating mario scenes in the latent space of a generative adversarial network. In *AAAI Conference on Artificial Intelligence*, 2021.

[14] Peter I. Frazier. A tutorial on Bayesian optimization. *arXiv:1807.02811 [cs, math, stat]*, 2018. arXiv: 1807.02811.

[15] Adam Gaier, Alexander Asteroth, and Jean-Baptiste Mouret. Aerodynamic design exploration through surrogate-assisted illumination. In *18th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, 2017.

[16] J. Guerrero-Viu, S. Hauns, S. Izquierdo, G. Miotto, S. Schrodi, A. Biedenkapp, T. Elsken, D. Deng, M Lindauer, and F. Hutter. Bag of baselines for multi-objective joint neural architecture search and hyperparameter optimization. arXiv:2105.01015 [cs.LG], 2021.

[17] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pages 507–523, 2011.

[18] Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[19] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[20] K. Kandasamy, W. Neiswanger, J. Schneider, B. Poczos, and E. Xing. Neural architecture search with Bayesian optimisation and optimal transport. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

[21] Paul Kent and Juergen Branke. BOP-Elites, a Bayesian optimisation algorithm for quality-diversity search. *arXiv:2005.04320 [cs, math, stat]*, 2020. arXiv: 2005.04320.

[22] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[23] Aaron Klein, Stefan Falkner, Jost Tobias Springenberg, and Frank Hutter. Learning curve prediction with Bayesian neural networks. In *International Conference on Learning Representations*, 2017.

[24] J. Knowles. ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 10(1):50–66, 2006.

[25] Joel Lehman and Kenneth O Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

[26] Joel Lehman and Kenneth O Stanley. Evolving a diversity of virtual creatures through novelty search and local competition. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, pages 211–218, 2011.

[27] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. In *International Conference on Learning Representations*, 2017.

[28] Edgar Liberis, Łukasz Dudziak, and Nicholas D Lane. μnas: Constrained neural architecture search for microcontrollers. In *Proceedings of the 1st Workshop on Machine Learning and Systems*, pages 70–79, 2021.

[29] Marius Lindauer, Katharina Eggensperger, Matthias Feurer, André Biedenkapp, Difan Deng, Carolin Benjamins, Tim Ruhopf, René Sass, and Frank Hutter. SMAC3: A versatile Bayesian optimization package for hyperparameter optimization. *arXiv:2109.09831 [cs, stat]*, 2022. arXiv: 2109.09831.

[30] H. Liu, K. Simonyan, O. Vinyals, C. Fernando, and K. Kavukcuoglu. Hierarchical representations for efficient architecture search. In *Proceedings of the International Conference on Learning Representations*, 2017.

[31] H. Liu, K. Simonyan, and Y. Yang. DARTS: Differentiable architecture search. In *Proceedings of the International Conference on Learning Representations*, 2019.

[32] Zhichao Lu, Kalyanmoy Deb, Erik Goodman, Wolfgang Banzhaf, and Vishnu Naresh Boddeti. NSGANetV2: Evolutionary multi-objective surrogate-assisted neural architecture search. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision*, pages 35—-51, 2020.

[33] Zhichao Lu, Whalen Ian, Vishnu Boddeti, Yashesh Dhebar, Kalyanmoy Deb, Erik Goodman, and Wolfgang Banzhaf. NSGA-net: A multi-objective genetic algorithm for neural architecture search, 2019.

[34] Bo Lyu, Hang Yuan, Longfei Lu, and Yunye Zhang. Resource-constrained neural architecture search on edge devices. *IEEE Transactions on Network Science and Engineering*, 9(1):134–142, 2021.

[35] H. Mendoza, A. Klein, M. Feurer, J. T. Springenberg, and F. Hutter. Towards automatically-tuned neural networks. In *ICML Workshop on Automatic Machine Learning*, 2016.

[36] Microsoft. Neural Network Intelligence, 2021.

[37] G. F. Miller, P. M. Todd, and S. U. Hegde. Designing neural networks using genetic algorithms. In *Proceedings of the Third International Conference on Genetic Algorithms*, pages 379–384, 1989.

[38] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv:1504.04909 [cs, q-bio]*, 2015. arXiv: 1504.04909.

[39] Michael Pearce and Juergen Branke. Continuous multi-task Bayesian optimisation with correlation. *European Journal of Operational Research*, 270(3):1074—-1085, 2018.

[40] Diego Perez-Liebana, Cristina Guerrero-Romero, Alexander Dockhorn, Dominik Jeurissen, and Linjie Xu. Generating diverse and competitive play-styles for strategy games. *arXiv preprint arXiv:2104.08641*, 2021.

[41] H. Pham, M. Guan, B. Zoph, Q. V. Le, and J. Dean. Efficient neural architecture search via parameter sharing. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4095–4104, 2018.

[42] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le. Regularized evolution for image classifier architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4780–4789, 2019.

[43] E. Real, S. Moore, A. Selle, S. Saxena, Y. L. Suematsu, J. Tan, Q. V. Le, and A. Kurakin. Large-scale evolution of image classifiers. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2902—-2911, 2017.

[44] R. Schmucker, M. Donini, V. Perrone, M. B. Zafar, and C. Archambeau. Multi-objective multi-fidelity hyperparameter optimization with application to fairness. In *NeurIPS Workshop on Meta-Learning*, 2020.

[45] Robin Schmucker, Michele Donini, Muhammad Bilal Zafar, David Salinas, and Cédric Archambeau. Multi-objective asynchronous successive halving. *arXiv:2106.12639 [cs, stat]*, 2021. arXiv: 2106.12639.

[46] Lennart Schneider, Florian Pfisterer, Martin Binder, and Bernd Bischl. Mutation is all you need. In *8th ICML Workshop on Automated Machine Learning*, 2021.

[47] Christoph Schorn, Thomas Elsken, Sebastian Vogel, Armin Runge, Andre Guntoro, and Gerd Ascheid. Automated design of error-resilient and hardware-efficient deep neural networks. *Neural Computing and Applications*, 32:18327−−18345, 2020.

[48] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, 2012.

[49] Dimitrios Stamoulis, Ruizhou Ding, Di Wang, Dimitrios Lymberopoulos, Bodhi Priyantha, Jie Liu, and Diana Marculescu. Single-Path NAS: Designing hardware-efficient ConvNets in less than 4 hours. In Ulf Brefeld, Elisa Fromont, Andreas Hotho, Arno Knobbe, Marloes Maathuis, and Céline Robardet, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 481–497, 2020.

[50] C. White, A. Zela, B. Ru, Y. Liu, and F. Hutter. How powerful are performance predictors in neural architecture search? In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[51] Colin White, Willie Neiswanger, Sam Nolen, and Yash Savani. A study on encodings for neural architecture search. In *Advances in Neural Information Processing Systems*, 2020.

[52] Colin White, Willie Neiswanger, and Yash Savani. Bananas: Bayesian optimization with neural architectures for neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[53] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10734–10742, 2019.

[54] C. Ying, A. Klein, E. Christiansen, E. Real, K. Murphy, and F. Hutter. NAS-Bench-101: Towards reproducible neural architecture search. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7105–7114, 2019.

[55] Yanqi Zhou, Siavash Ebrahimi, Sercan Ö Arık, Haonan Yu, Hairong Liu, and Greg Diamos. Resource-efficient neural architect. *arXiv preprint arXiv:1806.07912*, 2018.

[56] B. Zoph and Q. V. Le. Neural architecture search with reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2017.

[57] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8697–8710, 2018.