Neural Architecture Transfer

Zhichao Lu, Student Member, IEEE, Gautam Sreekumar, Erik Goodman, Wolfgang Banzhaf, Kalyanmoy Deb, Fellow, IEEE and Vishnu Naresh Boddeti, Member, IEEE

Abstract—Neural architecture search (NAS) has emerged as a promising avenue for automatically designing task-specific neural networks. Existing NAS approaches require one complete search for each deployment specification of hardware or objective. This is a computationally impractical endeavor given the potentially large number of application scenarios. In this paper, we propose Neural Architecture Transfer (NAT) to overcome this limitation. NAT is designed to efficiently generate task-specific custom models that are competitive under multiple conflicting objectives. To realize this goal we learn task-specific supernets from which specialized subnets can be sampled without any additional training. The key to our approach is an integrated online transfer learning and many-objective evolutionary search procedure. A pre-trained supernet is iteratively adapted while simultaneously searching for task-specific subnets. We demonstrate the efficacy of NAT on 11 benchmark image classification tasks ranging from large-scale multi-class to small-scale fine-grained datasets. In all cases, including ImageNet, NATnets improve upon the state-of-the-art under mobile settings (≤ 600M Multiply-Adds). Surprisingly, small-scale fine-grained datasets benefit the most from NAT. At the same time, the architecture search and transfer is orders of magnitude more efficient than existing NAS methods. Overall, experimental evaluation indicates that, across diverse image classification tasks and computational objectives, NAT is an appreciably more effective alternative to conventional transfer learning of fine-tuning weights of an existing network architecture learned on standard datasets. Code is available at https://github.com/human-analysis/neural-architecture-transfer.

Index Terms—Convolutional Neural Networks, Neural Architecture Search, AutoML, Transfer Learning, Evolutionary Algorithms.

1 INTRODUCTION

Image classification is a fundamental task in computer vision, where given a dataset and, possibly, multiple objectives to optimize, one seeks to learn a model to classify images. Solutions to address this problem fall into two categories: (a) Sufficient Data: A custom convolutional neural network architecture is designed and its parameters are trained from scratch using variants of stochastic gradient descent, and (b) Insufficient Data: An existing architecture designed on a large scale dataset, such as ImageNet [1], along with its pre-trained weights (e.g., VGG [2], ResNet [3]), is fine-tuned for the task at hand. These two approaches have emerged as the mainstays of present day computer vision.

Success of the aforementioned approaches is primarily attributed to architectural advances in convolutional neural networks. Initial efforts at designing neural architectures relied on human ingenuity. Steady advances by skilled practitioners has resulted in designs, such as AlexNet [4], VGG [2], GoogLeNet [5], ResNet [3], DenseNet [6] and many more, which have led to performance gains on the ImageNet Large Scale Visual Recognition Challenge [1]. In most other cases, a recent large scale study [7] has shown that, across many tasks, transfer learning by fine-tuning ImageNet pre-trained networks outperforms networks that are trained from scratch on the same data.

Moving beyond manually designed network architectures, Neural Architecture Search (NAS) [8] seeks to automate this process and find not only good architectures, but also their associated weights for a given image classification task. This goal has led to notable improvements in convolutional neural network architectures on standard image classification benchmarks, such as ImageNet, CIFAR-10 [9], CIFAR-100 [9] etc., in terms of predictive performance, computational complexity and model size. However, apart from transfer learning by fine-tuning the weights, current NAS approaches have failed to deliver new models for both weights and topology on custom non-standard datasets. The key barrier to realizing the full potential of NAS is the large data and computational requirements for employing existing NAS algorithms on new tasks.

In this paper, we introduce Neural Architecture Transfer (NAT) to breach this barrier. Given an image classification task, NAT obtains custom neural networks (both topology and weights), optimized for possibly many conflicting objectives, and does so without the steep computational burden of running NAS for each new task from scratch. A single run of NAT efficiently obtains multiple custom neural networks spanning the entire trade-off front of objectives.

Our solution builds upon the concept of a supernet [10] which comprises of many subnets. All subnets are trained simultaneously through weight sharing, and can be sampled very efficiently. This procedure decouples the network training and the search phases of NAS. A many-objective search can then be employed on top of the supernet to find all network architectures that provide the best trade-off among the objectives. However, training such supernets for each task from scratch is very computationally and data intensive. The key idea of NAT is to leverage an existing supernet and efficiently transfer it into a task-specific supernet, whilst simultaneously searching for architectures that offer the best trade-off between the objectives of interest. Therefore, unlike standard supernet-based NAS, we combine supernet transfer learning with the search process. At the conclusion of this process, NAT returns architectures having more than three objectives are called many-objective problems [11].
The key contributions of this paper are:

(i) subnets that span the entire objective trade-off front, and (ii) a task-specific supernet. The latter can now be utilized for all future deployment-specific NAS, i.e., new and different hardware or objectives, without any additional training.

The core of NAT’s efficiency lies in only adapting the subnets of the supernet that will lie on the efficient trade-off front of the new dataset, instead of all possible subnets. But, the structure of the corresponding subnets is unknown before adaptation. We resolve this “chicken-and-egg problem” by adopting an online procedure that alternates between the two primary stages of NAT: (a) supernet adaptation of subnets that are at the current trade-off front, and (b) evolutionary search for subnets that span the many-objective trade-off front. A pictorial overview of the entire NAT method is shown in Fig.1

In the adaptation stage, we first construct a layer-wise empirical distribution from the promising subnets returned by evolutionary search. Then, subnets sampled from this distribution are fine-tuned. In the search stage, to improve the efficiency of the search, we adopt a surrogate model to quickly predict the objectives of any sampled subnet without a full-blown and costly evaluation. Furthermore, the predictor model itself is also learned online from previously evaluated subnets. We alternate between these two stages until our computational budget is exhausted. The key contributions of this paper are:

- We introduce Neural Architecture Transfer as a NAS-powered alternative to fine-tuning based transfer learning. NAT is powered by a simple, yet highly effective online supernet fine-tuning and online accuracy predicting surrogate model.

- We demonstrate the scalability and practicality of NAT on multiple datasets corresponding to different scenarios: large-scale multi-class (ImageNet [1], CIFAR-10, CIFAR-100 [2]), medium-scale multi-class (CIFAR-10, CIFAR-100 [2]), small-scale multi-class (STL-10 [3]), large-scale fine-grained (Food-101 [4]), medium-scale fine-grained (Stanford Cars [5], FGVC Aircraft [6]) and small-scale fine-grained (DTD [7], Oxford-IIIT Pets [8], Oxford Flowers102 [9]) datasets.

- Under mobile settings (≤ 600M MAdds), NATNets lead to state-of-the-art performance across all these tasks. For instance, on ImageNet, NATNet achieves a Top-1 accuracy of 80.5% at 600M MAdds.

- We also demonstrate the utility of NAT in searching for a backbone for semantic segmentation, a dense prediction task. On Cityscapes [20], NAT matches the mIoU performance of Auto-DeepLab [21] while using 4× fewer MAdds.

- Finally, we demonstrate the scalability and utility of NAT across many objectives and on dense image prediction. Optimizing accuracy, model size and one of MAdds, CPU or GPU latency, NATNets dominate MobileNetV3 [22] across all objectives. We also consider a 12 objective problem of finding a common architecture across eleven datasets while minimizing MAdds. The best trade-off NATNet dominates all models across these datasets under mobile settings.

2 Related Work

Recent years have witnessed growing interests in neural architecture search. The promise of being able to automatically search for task-dependent network architectures is particularly appealing as deep neural networks are widely deployed in diverse applications and computational environments. Early methods [33], [34] made efforts to simultaneously evolve the topology of neural networks along with weights and hyperparameters. These methods perform competitively with hand-crafted networks on simple control tasks with shallow fully connected networks. Recent efforts [35] primarily focus on designing deep convolutional neural network architectures.

The development of NAS largely happened in two phases. Starting from NASNet [3], the focus of the first wave of methods was primarily on improving the predictive accuracy of CNNs including Block-QNN [36], Hierarchical NAS [37], and AmoebaNet [38], etc. These methods relied on Reinforcement Learning (RL) or Evolutionary Algorithm (EA) to search for an optimal modular structure that is repeatedly stacked together to form a network architecture. The search was typically carried out on relatively small-scale datasets (e.g. CIFAR-10/100 [9]), following which the best architectures were transferred to ImageNet for validation. A steady stream of improvements over state-of-the-art on numerous datasets were reported. The focus of the second wave of NAS methods was on improving the search efficiency.

A few methods have also been proposed to adapt NAS to other scenarios. These include meta-learning based approaches [39], [40] with application to few-shot learning tasks. XferNAS [41] and EAT-NAS [42] illustrate how architectures can be transferred between similar datasets or from smaller to larger datasets. Some approaches [43], [44] proposed RL-based NAS methods that...
### TABLE 1: Comparison of NAT and existing NAS methods. † indicates methods that scalarize multiple objectives into one composite objective or as an additional constraint, see text for details.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Search Method</th>
<th>Performance Prediction</th>
<th>Weight Sharing</th>
<th>Multiple Objective</th>
<th>Dataset Searched</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASNet [8]</td>
<td>RL</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>C10</td>
</tr>
<tr>
<td>PNAS [23]</td>
<td>SBMO</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>C10</td>
</tr>
<tr>
<td>DARTS [24]</td>
<td>gradient</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>C10</td>
</tr>
<tr>
<td>LEMONADE [25]</td>
<td>EA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>C10, C100, ImageNet64</td>
</tr>
<tr>
<td>ProxylessNAS [26]</td>
<td>RL / gradient</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ImageNet</td>
</tr>
<tr>
<td>MnasNet [27]</td>
<td>RL</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ImageNet</td>
</tr>
<tr>
<td>EfficientNet [28]</td>
<td>RL+scaling</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ImageNet</td>
</tr>
<tr>
<td>ChamNet [29]</td>
<td>EA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ImageNet</td>
</tr>
<tr>
<td>SPOS NAS [31]</td>
<td>EA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ImageNet</td>
</tr>
<tr>
<td>OnceForAll [32]</td>
<td>EA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ImageNet</td>
</tr>
<tr>
<td>FBNetV2 [33]</td>
<td>gradient</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ImageNet</td>
</tr>
<tr>
<td>NAT (this paper)</td>
<td>EA+transfer</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ImageNet, C10, C100, CSTIC-10, STL-10, Flowers102, Pets, DTD, Cars, Aircraft, Food-101</td>
</tr>
</tbody>
</table>

search on multiple tasks during training and transfer the learned search strategy, as opposed to searched networks, to new tasks at inference. Next, we provide short overviews on methods that are closely related to the technical approach in this paper. Table 1 provides a comparative overview of NAT to existing NAS approaches.

**Performance Prediction:** Evaluating the performance of an architecture requires a computationally intensive process of iteratively optimizing model weights. To alleviate this computational burden, regression models have been learned to predict an architecture’s performance without actually training it. Baker et al. [45] use a radial basis function to estimate the final accuracy of architectures from its accuracy in the first 25% of training iterations. PNAS [23] uses a multilayer perceptron (MLP) and a recurrent neural network to estimate the expected improvement in accuracy if the current modular structure (which is later stacked together to form a network) is expanded with a new branch. Conceptually, both of these methods seek to learn a prediction model that extrapolate (rather than interpolate), resulting in poor correlation in prediction. OnceForAll [31] also uses a MLP to predict accuracy from architecture encoding. However, the model is trained offline for the entire search space, thereby requiring a large number of samples for learning (16K samples -> 2 GPU-days for just constructing the surrogate model). Instead of using uniformly sampled architectures to train the prediction model to approximate the entire landscape, ChamNet [29] trains many architectures through full SGD and selects only 300 samples of high accuracy with diverse efficiency (Multiply-adds, Latency, Energy) to train a prediction model offline. In contrast, NAT learns a prediction model in an online fashion only on the samples at the current trade-off front as we explore the search space. Such an approach only needs to interpolate over a much smaller space of architectures constituting the current trade-off front. Consequently, this procedure significantly improves both the accuracy and the sample complexity of constructing the prediction model.

**Weight Sharing:** Approaches in this category involve training a supernet that contains all searchable architectures as its subnets. They can be broadly classified into two categories depending on whether the supernet training is coupled with architecture search or decoupled into a two-stage process. Approaches of the former kind [24, 26, 46] are computationally efficient but return suboptimal models. Numerous studies [47, 48, 49] allude to weak correlation between performance at the search and final evaluation stages. Methods of the latter kind [10, 31, 50] use performance of subnets (obtained by sampling the trained supernet) as a metric to select architectures during search. However, training a supernet beforehand for each new task is computationally prohibitive. In this work, we take an integrated approach where we train a supernet on large-scale datasets (e.g. ImageNet) once and couple it with our architecture search to quickly adapt it to a new task. An elaborated discussion connecting our method to existing approaches is provided in the supplementary materials.

**Multi-Objective NAS:** Methods that consider multiple objectives for designing hardware specific models have also been developed. The objectives are optimized either through (i) scalarization, or (ii) Pareto-based solutions. The former include, ProxylessNAS [26], MnasNet [27], ChamNet [29], MobileNetV3 [32], and FBNetV2 [32] which use a scalarized objective or an additional constraint to encourage high accuracy and penalize compute inefficiency at the same time, e.g., maximize $\text{Acc} \cdot (\text{Latency/Target})^{-0.07}$. Conceptually, the search of architectures is still guided by a single objective and only one architecture is obtained per search. Empirically, multiple runs with different weighting of the objectives are needed to find an architecture with the desired trade-off, or multiple architectures with different complexities. Methods in the latter category include [24, 31, 52, 53, 54] and aim to approximate the entire Pareto-efficient frontier simultaneously—i.e. multiple architectures with different complexities are obtained in a single run. These approaches rely on heuristics (e.g., EA) to efficiently navigate the search space allowing practitioners to visualize the trade-off between the objectives and to choose a suitable network *a posteriori* to the search. NAT falls into the latter category and uses an accuracy prediction model and weight sharing for efficient architecture transfer to new tasks.

## 3 Proposed Approach

**Neural Architecture Transfer** consists of three main components: an accuracy predictor, an evolutionary search routine, and a supernet. NAT starts with an archive $\mathcal{A}$ of architectures (subnets) created by uniform sampling from our search space. We evaluate the performance $f_i$ of each subnet ($\alpha_i$) using weights inherited from the supernet. The accuracy predictor is then constructed from $(\alpha_i, f_i)$ pairs which (jointly with any additional objectives provided by the user) drives the subsequent many-objective evolutionary search towards optimal architectures. Promising architec-
tures at the conclusion of the evolutionary process are added to the archive \( \mathcal{A} \). The (partial) weights of the supernet corresponding to the top-ranked subnets in the archive are fine-tuned. NAT repeats this process for a pre-specified number of iterations. At the conclusion, we output both the archive and the task-specific supernet. Networks that offer the best trade-off among the objectives can be post-selected from the archive. Detailed descriptions of each component of NAT are provided in the following subsections.

Figure 1 and Algorithm 1 provide an overview of our entire approach.

**Algorithm 1: Neural Architecture Transfer**

**Input**: Training data \( D_{trn} \), validation data \( D_{old} \), additional objectives \( f \), supernet \( S_w \), archive size \( N \), # of iterations \( T \), # of epochs \( E \), # of generations \( G \).

1. \( t \leftarrow 0 \) // initialize an iteration counter.
2. \( A \leftarrow \) randomly initialize an archive of archs with a size of \( N \).
3. \( \text{while} t < T \) do
   4. \( f \leftarrow S_w(A, D_{old}) \) // compute accuracy by inheriting weights and inference.
   5. \( f \leftarrow \) Accuray Predictor \((A, f) \) \( \triangleleft \) Algo. 2
   6. \( P_t \leftarrow \) Evolutionary Search \((S_f, f, A, G) \) \( \triangleleft \) Algo. 3
   7. \( \text{if} \) promising archs by evolutionary search.
   8. \( \text{if} \) keep the top-\( N \) ranked archs in archive.
   9. \( A \leftarrow \text{Selection}(A \cup P_t, N) \) \( \triangleleft \) Algo. 4
   10. \( S_w \leftarrow \text{Adapt}(S_w, A, D_{trn}, E) \) \( \triangleleft \) Algo. 5
   11. \( t \leftarrow t + 1 \) // optional case of no preferences from users.
12. \( A^* \leftarrow \) choose a subset of archs from \( A \) based on trade-offs by method presented in supplementary materials under Section 3.
13. **Return** \( S_w, A, A^* \).

### 3.1 Problem Formulation

The problem of neural architecture search for a target dataset \( D = \{D_{trn}, D_{old}, D_{lst}\} \) with many objectives can be formulated as the following bilevel optimization problem [55].

\[
\begin{align*}
\text{minimize} & \quad F(a) = (f_1(a; w^*(a)), \ldots, f_m(a; w^*(a)))^T, \\
\text{subject to} & \quad w^*(a) \in \arg \min \mathcal{L}(w; a), \\
& \quad a \in \Omega_a, \quad w \in \Omega_w,
\end{align*}
\]

where the upper-level variable \( a \) defines a candidate architecture, and the lower-level variable \( w(a) \) denotes its associated weights. \( \mathcal{L}(w; a) \) is the cross-entropy loss on the training data \( D_{trn} \) for an architecture \( a \). \( F : \Omega \rightarrow \mathbb{R}^m \) constitutes \( m \) (user-) desired, possibly competing, objectives—e.g., predictive performance on validation data \( D_{old} \), number of parameters (\#Params), multiply-adds (\#MAdds), latency / power consumption / memory footprint on specific hardware etc.

The bi-level optimization is typically solved in an iterative fashion, with an inner optimization loop over the weights of the network for a given architecture, and an outer optimization loop over the network architectures themselves. The computational challenge of solving this problem stems from both the upper and lower level optimization. Learning optimal weights of a network in the lower level necessitates costly iterations of stochastic gradient descent over multiple epochs. Similarly, searching the optimal architecture on the upper level is prohibitive due to the discrete nature of the architecture description, size of search space and our desire to optimize many, possibly conflicting, objectives.

![Fig. 2: The architectures in our search space are variants of MobileNetV2 family of models [22, 27, 28, 56]. (a) Each network consists of five stages. Each stage has two to four layers. Each layer is an inverted residual bottleneck structure [56].](image)

### 3.2 Search Space and Encoding

The search for optimal network architectures can be performed over many different search spaces. The generality of the chosen search space has a major influence on the quality of results that are feasible. We adopt a modular design for overall structure of the network, consisting of a stem, multiple stages and a tail (see Fig. 2a). The stem and tail are common to all networks and not searched. Each stage in turn comprises of multiple layers, and each layer itself is an inverted residual bottleneck structure [56].

- **Network**: We search for the input image resolution and the width multiplier (\( W \), the number of layers in each stage, the \( \# \) of output channels (expansion ratio \( E \)) of the first \( 1 \times 1 \) convolution and the kernel size \( K \) of the depth-wise separable convolution in each layer. 

  (b) **Encoding**: We segment the CNN architecture into five sequentially connected stages. The stages gradually reduce the feature map size and increase the number of channels (Fig. 2a Middle).

- **Layer**: We search over the expansion ratio (between the \( \# \) of output and input channels) of the first \( 1 \times 1 \) convolution and the kernel size of the depth-wise separable convolution (Fig. 2a Right).
Overall, we search over four primary hyperparameters of CNNs i.e., the depth (# of layers), the width (# of channels), the kernel size, and the input resolution. The resulting volume of our search space is approximately $3.5 \times 10^{19}$ for each combination of image resolution and width multiplier.

To encode these architectural choices, we use an integer string of length 22, as shown in Fig. 2. The first two values represent the input image resolution and width multiplier, respectively. The remaining 20 values denote the expansion ratio and kernel size settings for each of the 20 layers. The available options for expansion ratio and kernel size are [3, 4, 6] and [3, 5, 7], respectively. It is worth noting that we sort the layer settings in ascending #MAdds order, which is beneficial to the mutation operator used in our evolutionary search algorithm.

### 3.3 Accuracy Predictor

The main computational bottleneck of NAS arises from the nested nature of the bi-level optimization problem. The inner optimization requires the weights of the subnets to be thoroughly learned prior to evaluating its performance. Methods like weight-sharing [21], [46], [50] allow sampled subnets to inherit weights among themselves or from a supernet, avoiding the time-consuming process (typically requiring hours) of learning weights through SGD. However, standalone weight-sharing still requires inference on validation data (typically requiring minutes) to assess performance. Therefore, simply having to evaluate the subnets can still render the overall process computationally prohibitive for methods [8], [27], [38] that sample thousands of architectures during search.

To mitigate the computational burden of fully evaluating the subnets, we adopt a surrogate accuracy predictor that regresses the performance of a sampled subnet without performing training or inference. By learning a functional relation between the integer-strings (subnets in the encoded space) and the corresponding performance, this approach decouples the evaluation of an architecture from data-processing (including both SGD and inference). Consequently, the evaluation time reduces from hours/minutes to seconds. We illustrate this concept in Fig. 3. The effectiveness of this idea, however, is critically dependent on the quality of the surrogate model. Below we identify three desired properties of such a model:

1. Reliable prediction: high rank-order correlation between predicted and true performance.
2. Consistent prediction: the quality of the prediction should be consistent across different datasets.
3. Sample efficiency: minimizing the number of training examples necessary to construct an accurate predictor model, since each training sample requires costly training and evaluation of a subnet.

Current approaches [23], [29], [31] that use surrogate based accuracy predictors, however, do not satisfy property (1) and (3) simultaneously. For instance, PNAS [23] uses 1,160 subnets to build the surrogate but only achieves a rank-order correlation of 0.476. Similarly, OnceForAll [31] uses 16,000 subnets to build the surrogate. The poor sample complexity and rank-order correlation of these approaches, is due to the offline learning of the surrogate model. Instead of focusing on models that are at the trade-off front of the objectives, these surrogate models are built for the entire search space. Consequently, these methods require a significantly larger and more complex surrogate model.

We overcome the aforementioned limitation by restricting the surrogate model to the search space that constitutes the current objective trade-off. Such a solution significantly reduces the sample complexity of the surrogate and increases the reliability of its predictions. We adopt four low-complexity predictors, namely, Gaussian Process (GP) [29], Radial Basis Function (RBF) [45], Multilayer Perceptron (MLP) [23], and Decision Tree (DT) [58]. Empirically, we observe that RBFs are consistently better than the other three models if the # of training samples is more than 100. To further improve RBF’s performance, especially under a high sample efficiency regime, we construct an ensemble of RBF models. As outlined in Algorithm 2, each RBF model is constructed with a subset of samples and features randomly selected from the training instances. The correlation between predicted accuracy and true accuracy from an ensemble of 500 RBF models outperforms all

---

**Algorithm 2: Accuracy Predictor (RBF Ensemble)**

```python
Input : Training data $X$, training targets $Y$, ensemble size $K$
1. $k \leftarrow 0$ // initialize a counter.
2. pool $\leftarrow \emptyset$ // initialize a pool to store all models.
3. while $k < K$ do
4.   $(\hat{X}, \hat{Y}) \leftarrow$ randomly create a subset of the training data.
5.   idx $\leftarrow$ randomly pick a subset of the features in training data.
6.   rbf $\leftarrow$ fit a RBF model from $\hat{X}[:, \text{idx}]$ and $\hat{Y}$.
7.   pool $\leftarrow$ pool $\cup$ (rbf, idxs) // append the fitted model to the pool.
8.   $k \leftarrow k + 1$
9. end
10. Return a pool of $K$ RBF models.
```

---

**Fig. 4:** Accuracy predictor performance as a function of training samples. For each model, we show the mean and standard deviation of the Spearman rank correlation on 11 datasets (Table 3). The size of RBF ensemble is 500.
other models across all regimes. Fig. 4 compares the performance of the different surrogate models we considered. Practically, we observed that the RBF ensemble can be learned under a minute.

### 3.4 Many-Objective Evolutionary Search

Given the accuracy predictor, we employ a customized evolutionary algorithm (EA) to search for optimal architectures that offer the best trade-off between many objectives. The EA is an iterative process in which initial architectures, selected from the archive of previously explored architectures, are gradually improved as a group, referred to as a population. In every generation (iteration), a group of offspring (i.e., new architectures) are created by applying variations through crossover and mutation (described below) operations on the most promising architectures, also known as parents, found so far in the population. Every member of the population, i.e., both parents and offspring, competes for survival and reproduction (becoming a parent) in each generation. See Fig. 1 (bottom right shaded in green) for a pictorial overview, and Algorithm 3 for the pseudocode.

**Crossover** exchanges information between two or more population members to create two or more new members. Designing an effective crossover between non-standard solution representations can be difficult and has been largely ignored by existing EA-based NAS algorithms [37], [38], [59]. Here we adopt a customized, homogeneous crossover that uniformly picks integers from parent architectures to create offspring architectures. This crossover operator offers two properties: (1) it preserves common integers shared between parents; and (2) it is free of additional hyperparameters. Fig. 5a visualizes our implementation of the crossover operation. We generate two offspring architectures with each crossover, and an offspring population of the same size as the parent population is generated in each generation.

**Mutation** is a local operator that perturbs a solution to produce a new solution in its vicinity. In this work, we use a discretized version of the polynomial mutation (PM) operator [60] and apply it to every solution created by the crossover operator. For a given architecture $\mathbf{a}$, PM is carried out integer-wise with probability $p_m$.

![Algorithm 3: Evolutionary Search](image)

```
Algorithm 3: Evolutionary Search

Input : Accuracy predictor $S_f$, additional objectives $\tilde{f}$, archive of arches $\mathcal{A}$, max. # of generations $G$, population size $K$, crossover probability $p_c$, mutation probability $p_m$.
1: $g \leftarrow 0$ // initialize an generation counter.
2: $f \leftarrow S_f(\mathcal{A})$ // compute accuracy of all arches in archive.
3: $P \leftarrow \text{Selection}(\mathcal{A}, f, \tilde{f}(\mathcal{A}), K)$ // initialize the parent population with top-$K$ ranked arches from $\mathcal{A}$.
4: while $g < G$ do
5: // choose parents through tournament selection for mating.
6: $P \leftarrow \text{Binary Tournament Selection}(P)$
7: // create offspring population by crossover between parents.
8: $Q \leftarrow \text{Crossover}(P, p_c)$
9: // induce randomness to offspring population through mutation.
10: $Q \leftarrow \text{Mutation}(Q, p_m)$
11: $R \leftarrow P \cup Q$ // merge parent and offspring population.
12: // survive the top-$K$ archs to next generation.
13: $P \leftarrow \text{Selection}(R, S_f(R), \tilde{f}(R), K)$
14: $g \leftarrow g + 1$
15: end
16: Return parent population $P$.
```

and the mutated $i^{th}$ integer, $\mathbf{a}_i$, of the mutated offspring is:

$$
\mathbf{a}_i^t = \begin{cases} 
\mathbf{a}_i + \frac{((2u)^{1/(1+p_m)} - 1)(\mathbf{a}_i - \mathbf{a}_i^{(L)})}{1/(1+p_m)}, & \text{for } u \leq 0.5, \\
\mathbf{a}_i + (1 - (2(1-u))^{1/(1+p_m)})(\mathbf{a}_i^{(U)} - \mathbf{a}_i), & \text{for } u > 0.5 
\end{cases}
$$

where $u$ is a uniform random number in the interval $[0, 1]$. $\mathbf{a}_i^{(L)}$ and $\mathbf{a}_i^{(U)}$ are the lower and upper bounds of $\mathbf{a}_i$, respectively. Each mutated value in an offspring is rounded to the nearest integer. The PM operator inherits the parent-centric convention, in which the offspring are intentionally created around the parents. The centricity is controlled via an index hyperparameter $\eta_m$. In particular, high-values of $\eta_m$ tend to create mutated offspring around the parent, and low-values encourage mutated offspring to be further away from the parent architecture. See Fig. 5b for a visualization of the effect of $\eta_m$. It is worth noting that the PM operator was originally proposed for continuous optimization where distances between variable values are naturally defined. In contrast, in context of our encoding, our variables are categorical in nature, indicating a particular layer hyperparameter. So we sort the searched subnets in ascending order of #MAdds, such that $\eta_m$ now controls the difference in #MAdds between the parent and the mutated offspring.

We apply PM to every member in the offspring population (created from crossover). We then merge the mutated offspring population with the parent population and select the top half using many-objective selection operator described in Algorithm 3. This procedure creates the parent population for the next generation. We repeat this overall process for a pre-specified number of generations and output the parent population at the conclusion of the evolution.

### 3.5 Many-Objective Selection

In addition to high predictive accuracy, real-world applications demand NAS algorithms to simultaneously balance a few other conflicting objectives that are specific to the deployment scenarios. For instance, mobile or embedded devices often have restrictions in terms of model size, multiply-adds, latency, power consumption, and memory footprint. With no prior assumption on the correlation among these objectives, a scalable (to the number of objectives) selection is required to drive the search towards the high dimensional Pareto front. In this work, we adopt the reference point guided selection originally proposed in NSGA-III [11], which has been shown to be effective in handling problems...
Algorithm 4: Reference Point Based Selection

Input: A set of archs \( R \), their objectives \( F \), number of archs to select \( N \), reference directions \( Z \).

1. put archs into different fronts (rank levels) based on domination.
2. \( (F_1, F_2, \ldots) \leftarrow \text{non-dominated sort}(F) \)
3. \( S \leftarrow \emptyset, i \leftarrow 1 \)
4. while \(|S| + |F_i| < N\) do \( S \leftarrow S \cup F_i; i \leftarrow i + 1 \);
5. \( F_L \leftarrow F_i \) // next front is the split front where we cannot accommodate all archs associated with it.
6. if \(|S| + |F_L| = N\) then \( S \leftarrow S \cup F_L \);
7. else
8. \( (\tilde{S}, \tilde{F}_L) \leftarrow \text{Normalize}(S, F_L) \) // normalize the objectives based on ideal and nadir points derived from \( R \).
9. \( d \leftarrow \text{compute orthogonal dist to } Z_i \) for each \( i \)
10. \( \rho \leftarrow \text{count associated solutions for } Z_i \) based on \( d \) for each \( i \).
11. // remaining archs from \( F_L \) to fill up \( S \).
12. \( S \leftarrow S \cup \text{Niching}(\tilde{F}_L, N - |S|, \rho, d) \)
13. end
14. Return \( S \).

Fig. 6: (a) An example (assuming minimization of all objectives) of the selection process in Alg. 4. We first create reference directions \( Z \) by joining reference points with the ideal solution (origin). Then through \textit{non-dominated sort}, three non-dominated solutions are identified, associated with reference directions \( Z^{(1)}, Z^{(2)} \) and \( Z^{(3)} \). We then select the remaining solutions by the orthogonal distances to the reference directions with no associated solutions—i.e., \( Z^{(2)} \) and \( Z^{(3)} \). This selection is scalable to larger \# of objectives. A tri-objective example is shown in (b).

with as many as 15 objectives. In the remainder of this section, we provide an overview of NSGA-III procedure and refer readers to the original publication for more details.

**Domination** is a widely-used partial ordering concept for comparing two objective vectors. For a generic many-objective optimization problem: \( \min_{a} \{ f_1(a), \ldots, f_m(a) \} \), where \( f_i(.) \) are the objectives (say, loss functions) to be optimized and \( a \) is the representation of a neural network architecture. For two given solutions \( a_1 \) and \( a_2 \), solution \( a_1 \) is said to dominate \( a_2 \) (i.e., \( a_1 \preceq a_2 \)) if following conditions are satisfied:

1. \( a_1 \) is no worse than \( a_2 \) for all objectives \( f_i(a_1) \leq f_i(a_2), \forall i \in \{0, \ldots, m\} \), and
2. \( a_1 \) is strictly better than \( a_2 \) in at least one objective \( \exists i \in \{0, \ldots, m\} \) such that \( f_i(a_1) < f_i(a_2) \).

A solution \( a_2 \) is said to be non-dominated if these conditions hold against all the other solutions \( a_j \) (with \( j \neq i \)) in the entire search space of \( a \).

With the above definition, we can sort solutions to different ranks of domination, where solutions in the same rank are non-dominated to each other, and there exists at least one solution in lower rank that dominates any solution in the higher rank. Thus, a lower non-dominated ranked set is lexicographically better than a higher ranked set. This process is referred as \textit{non-dominated sort}, and it is the first step in the selection process. During the many-objective selection process, the lower ranked sets are chosen one at a time until no more sets can be included to maintain the population size. The final accepted set may have to be \textit{split} to choose only a part. For this purpose, we choose the most diverse subset based on a diversity-maintaining mechanism. We first create a set of reference directions from a set of uniformly distributed \((m - 1)\)-dimensional space reference points in the unit simplex by using Das-and-Dennis method [61]. Then we associate each solution to a reference direction based on orthogonal distance of the solution from the direction. Then, for every reference direction, we choose the closest associated solution in a systematic manner by adaptively computing a niche count \( \rho \) so that every reference direction gets an equal opportunity to choose a representative closest solution in the selected population. The domination and diversity-preserving procedures are easily scalable to any number of objectives and importantly are free from any user-defined hyperparameter. See Algorithm 4 for the pseudocode and Fig. 6 for a graphical illustration. A more elaborated discussion on the necessity of the reference point based selection is provided in the supplementary materials under Section 2.

### 3.6 SuperNet Adaptation

Instead of training every architectures sampled during search from scratch, NAS with weight sharing [24], [46] inherits weights from previously-trained networks or from a supernet. Directly inheriting the weights obviates the need to optimize the weights from scratch and speeds up the search from thousands of GPU days to only a few. In this work, we focus on the supernet approach [10], [31]. It involves first training a large network model (in which searchable architectures become subnets) prior to the search. Then the performance of the subnets, evaluated with the inherited weights, is used to guide the selection of architectures during search. The key to the success of this approach is that the performance of the subnets with the inherited weights be highly correlated with the performance of the same subnet when thoroughly trained from scratch. Satisfying this desideratum necessitates that the supernet weights be learned in such a way that all subnets are optimized simultaneously.

Existing methods [30], [53] attempt to achieve the above goal by imposing \textit{fairness} in training the supernet, where the probabilities of training any particular subnet for each batch of data is uniform in expectation. However, we argue that simultaneously training all the subnets in the search space is practically not feasible and, more importantly, not necessary. Firstly, it is evident from existing NAS approaches [26], [62] that different objectives (#Params, #MAdds, latency on different hardware, etc.) require different architectures in order to be efficient. In other words, not all subnets are equally important for the task at hand. Secondly, only a tiny fraction of the search space can practically be explored by a NAS algorithms.

Based on the aforementioned observations, we propose a simple yet effective supernet training routine that only focuses on training the subnets recommended by the evolutionary search algorithm in Section 3.5. Specifically, we seek to exploit the knowledge gained from the search process so far. Recall that our algorithm uses an archive to keep track of the promising architectures explored so far. For each value in our architecture encoding, we construct a categorical distribution from architectures in the
Algorithm 5: Adapt SuperNet

\begin{algorithm}
\begin{algorithmic}
\State \textbf{Input :} Supernet $S_{\text{sup}}$, archive of archs $A$, training data $D_{\text{trn}}$.
\State $e \leftarrow 0$ // initialize an epoch counter.
\While {$e < E$}
\For {each batch in $D_{\text{trn}}$}
\State $\text{subnet} \leftarrow$ sample from $\text{Distr}$.
\State $w \leftarrow$ set forward path of $S_{\text{w}}$ according to $\text{subnet}$.
\State $\mathcal{L} \leftarrow$ compute cross-entropy loss on data batch.
\State $\nabla w \leftarrow$ compute the gradient by $\partial \mathcal{L} / \partial w$
\State $S_{\text{w}} \leftarrow$ one step of SGD.
\EndFor
\State $e \leftarrow e + 1$
\EndWhile
\State \textbf{Return} supernet $S_{\text{w}}$.
\end{algorithmic}
\end{algorithm}

archive, where the probability for $i^{th}$ integer taking on the $j$ value is computed as:

$$p(X_i = j) = \frac{\# \text{ of architectures with option } j \text{ at } i^{th} \text{ integer}}{\text{total } \# \text{ of architectures in the archive}}$$ (3)

In each training step (batch of data), we sample an integer-string from the above distribution. We then activate the subparts of the supernet corresponding to the architecture decoded from the integer-string. Only weights corresponding to the activated subparts in the supernet will be updated in each step. See Algorithm 5 for pseudocode. A more in-depth discussion connecting our proposed approach to the existing supernet-based NAS approaches is provided in the supplementary materials under Section 1.

4 Experimental Evaluation

In this section, we present experimental results to evaluate the efficacy of Neural Architecture Transfer on multiple image classification tasks. In addition, we also investigate the scalability of our approach to more than two objectives. For all the experiments in this section, we use the same set of hyperparameters (see Table 2) for the different components of NAT. These choices were guided by the ablation studies described in Section 5.

4.1 Datasets

We consider eleven image classification datasets for evaluation with sample size varying from 2,040 to 180,000 images (20 to 18,000 images per class; Table 3). These datasets span a wide variety of image classification tasks, including superordinate-level recognition (ImageNet [1], CIFAR-10 [2], CIFAR-100 [2], CINIC-10 [12], STL-10 [13]); fine-grained recognition (Food-101 [14], Oxford Flowers102 [19]); and texture classification (DTD [17]). We use the ImageNet dataset for training the supernet, and use the other ten datasets for architecture transfer.

4.2 Supernet Preparation

Our supernet is constructed by setting the architecture encoding at the maximum value, i.e., four layers in each stage and every layer uses expand ratio of six and kernel size of seven. Adapting subnets of a supernet with randomly initialized weights leads to training instability and large variance in its performance. Therefore, we warm-up the supernet weights on ImageNet following the progressive shrinking algorithm [31], where the supernet is first trained at full-scale, with subnets corresponding to different options (expand ratio, kernel size, # of layers) being gradually activated during the training process. This procedure, which takes about 6 days on a server with eight V100 GPUs, is optimized with only the cross-entropy loss i.e., a single objective. We note that supernet preparation expense is a one-time cost that amortizes over any subsequent transfer to different datasets and objective combinations we show in the following subsections.

4.3 ImageNet Classification

Before we evaluate our approach for architecture transfer to other datasets, we first validate its effectiveness on the ImageNet-1K dataset. This experiment evaluates the effectiveness of NAT in adapting and searching for architectures that span trade-off between two objectives. For this experiment, we consider accuracy and #MAdds as the two objective of interest. We randomly sample 50,000 images from the original ImageNet training set as the validation set to guide the architecture search. We run NAT for 30 iterations, and from the final archive of architectures, we select four models ranging from 200M MAdds to 600M MAdds (for high-end mobile devices). Following [31], we fine-tune\textsuperscript{5} each model to further boost the performance. Our fine-tune training largely follows [27]: RMSProp optimizer with decay 0.9 and momentum 0.9; batch normalization momentum 0.99; weight decay 1e-5. We use a batch size of 512 and an initial learning rate of 0.012 that gradually reduces to zero following the cosine annealing schedule. Our regularization settings are similar as in [28]: we use augmentation policy [63], drop connect ratio 0.2, and dropout rate 0.2.

Table 3 shows the performance of NAT models obtained through bi-objective optimization of maximizing accuracy and minimizing #MAdds. Our models, referred to as NAT-M\{1,2,3,4\}, are in ascending order of #MAdds (Fig. 4). Fig. 5 shows the full

| Table 2: Hyperparameter Settings |
|-----------------|-----------------|-----------------|
| Category        | Parameter       | Setting         |
| Global          | Archive size    | 300             |
|                 | Number of iterations | 30              |
| Accuracy predictor | Train size     | 100             |
|                 | Ensemble size   | 500             |
| Evolutionary search | Population size | 100             |
|                 | Number of generations per iteration | 100          |
|                 | Crossover probability | 0.9           |
|                 | Mutation probability | 0.1          |
|                 | Mutation index $\tau_m$ | 1.0          |
| Supernet        | Number of epochs per iteration | 5            |

Stanford Cars [15], FGVC Aircraft [16], Oxford-IIIT Pets [18], Oxford Flowers102 [19]); and texture classification (DTD [17]). We use the ImageNet dataset for training the supernet, and use the other ten datasets for architecture transfer.

6. Section 5.5 studies the impact of this fine-tuning step.

| Table 3: Benchmark Datasets for Evaluation |
|-----------------|-----------------|-----------------|
| Dataset         | Type            | Train Size      | Test Size     | #Classes |
| CIFAR-10        | multi-class     | 180,000         | 9,000         | 10       |
| CIFAR-100       | multi-class     | 50,000          | 10,000        | 10       |
| STL-10          | multi-class     | 5,000           | 8,000         | 10       |
| STL-10          | fine-grained    | 6,667           | 3,333          | 100      |
| DTD             | fine-grained    | 3,760           | 1,880          | 47       |
| Oxford-IIIT Pets | fine-grained    | 3,680           | 3,369          | 37       |
| Oxford Flowers102 | fine-grained   | 2,040           | 6,149          | 102      |
| CINIC-10        | multi-class     | 1,281,167       | 50,000         | 1,000    |
| ImageNet        | multi-class     | 1,281,167       | 50,000         | 1,000    |
| CIFAR-10        | multi-class     | 180,000         | 9,000         | 10       |
| CIFAR-100       | multi-class     | 50,000          | 10,000        | 10       |
| CIFAR-100       | fine-grained    | 50,000          | 10,000        | 10       |
| STL-10          | multi-class     | 5,000           | 8,000         | 10       |
| STL-10          | fine-grained    | 6,667           | 3,333          | 100      |
| DTD             | fine-grained    | 3,760           | 1,880          | 47       |
| Oxford-IIIT Pets | fine-grained    | 3,680           | 3,369          | 37       |
| Oxford Flowers102 | fine-grained   | 2,040           | 6,149          | 102      |

5. A visualization of such distributions is shown in Fig.3 of supplementary material.
### TABLE 4: ImageNet-1K Classification

NATNets comparison with manual and automated design of efficient convolutional neural networks. Models are grouped into sections for better visualization. Our results are underlined and the best result in each section is in bold. CPU latency (batchsize=1) is measured on Intel i7-8700K and GPU latency (batchsize=64) is measured on 1080Ti. “WS” stands for weight sharing. All methods are under single crop and single model condition, without any additional data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>#Params</th>
<th>#Multi-Adds</th>
<th>CPU Lat (ms)</th>
<th>GPU Lat (ms)</th>
<th>Top-1 Acc (%)</th>
<th>Top-5 Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAT-M1</td>
<td>WS+EA</td>
<td>6.0M</td>
<td>225M</td>
<td>9.1</td>
<td>30</td>
<td>77.5</td>
<td>93.5</td>
</tr>
<tr>
<td>MobileNetV2 [56]</td>
<td>manual</td>
<td>3.5M</td>
<td>300M</td>
<td>8.3</td>
<td>23</td>
<td>72.0</td>
<td>91.0</td>
</tr>
<tr>
<td>SPOS NAS [35]</td>
<td>RS+EA</td>
<td>3.4M</td>
<td>328M</td>
<td>-</td>
<td>-</td>
<td>74.7</td>
<td>92.0</td>
</tr>
<tr>
<td>ProxylessNAS [26]</td>
<td>RL/gradient</td>
<td>4.0M</td>
<td>465M</td>
<td>8.5</td>
<td>27</td>
<td>75.1</td>
<td>92.5</td>
</tr>
<tr>
<td>MnasNet-A1 [27]</td>
<td>RL</td>
<td>3.9M</td>
<td>312M</td>
<td>9.3</td>
<td>31</td>
<td>75.2</td>
<td>92.5</td>
</tr>
<tr>
<td>MobileNetV3 [22]</td>
<td>RL/NetAdapt</td>
<td>5.4M</td>
<td>219M</td>
<td>10.6</td>
<td>33</td>
<td>75.2</td>
<td>-</td>
</tr>
<tr>
<td>MUXNet-m [54]</td>
<td>EA</td>
<td>3.4M</td>
<td>218M</td>
<td>14.7</td>
<td>42</td>
<td>73.5</td>
<td>92.5</td>
</tr>
<tr>
<td>FBNetV2-F4 [52]</td>
<td>gradient</td>
<td>7.0M</td>
<td>238M</td>
<td>15.6</td>
<td>44</td>
<td>76.0</td>
<td>-</td>
</tr>
<tr>
<td>NAT-M2</td>
<td>WS+EA</td>
<td>7.7M</td>
<td>312M</td>
<td>11.4</td>
<td>37</td>
<td>78.6</td>
<td>94.3</td>
</tr>
<tr>
<td>MUXNet-1 [54]</td>
<td>EA</td>
<td>4.0M</td>
<td>318M</td>
<td>19.2</td>
<td>74</td>
<td>76.6</td>
<td>93.2</td>
</tr>
<tr>
<td>EfficientNet-B0 [30]</td>
<td>RL/scaling</td>
<td>5.3M</td>
<td>390M</td>
<td>14.4</td>
<td>46</td>
<td>77.1</td>
<td>93.2</td>
</tr>
<tr>
<td>AtomNAS-C+ [65]</td>
<td>WS+shrinkage</td>
<td>5.9M</td>
<td>363M</td>
<td>-</td>
<td>-</td>
<td>77.6</td>
<td>93.5</td>
</tr>
<tr>
<td>AutoNL-L [66]</td>
<td>gradient</td>
<td>5.6M</td>
<td>353M</td>
<td>-</td>
<td>-</td>
<td>77.7</td>
<td>93.7</td>
</tr>
<tr>
<td>DNA-c [66]</td>
<td>gradient</td>
<td>5.3M</td>
<td>466M</td>
<td>14.5</td>
<td>67</td>
<td>77.8</td>
<td>93.7</td>
</tr>
<tr>
<td>NAT-M3</td>
<td>ResNet-152 [4]</td>
<td>manual</td>
<td>60M</td>
<td>11.3B</td>
<td>176</td>
<td>77.8</td>
<td>93.8</td>
</tr>
<tr>
<td>MixNet-L [47]</td>
<td>RL</td>
<td>7.3M</td>
<td>565M</td>
<td>29.4</td>
<td>105</td>
<td>78.9</td>
<td>94.2</td>
</tr>
<tr>
<td>EfficientNet-B1 [28]</td>
<td>RL/scaling</td>
<td>7.8M</td>
<td>700M</td>
<td>19.5</td>
<td>67</td>
<td>79.1</td>
<td>94.4</td>
</tr>
<tr>
<td>NAT-M4</td>
<td>WS+EA</td>
<td>9.1M</td>
<td>490M</td>
<td>16.1</td>
<td>62</td>
<td>79.9</td>
<td>94.9</td>
</tr>
<tr>
<td>BigNASModel-L [68]</td>
<td>WS</td>
<td>6.4M</td>
<td>318M</td>
<td>19.2</td>
<td>74</td>
<td>76.6</td>
<td>93.2</td>
</tr>
<tr>
<td>OnceForAll [31]</td>
<td>WS+EA</td>
<td>9.1M</td>
<td>13B</td>
<td>486.6</td>
<td>206</td>
<td>80.0</td>
<td>94.9</td>
</tr>
<tr>
<td>Inception-v4 [67]</td>
<td>manual</td>
<td>48M</td>
<td>13B</td>
<td>99.1</td>
<td>289</td>
<td>80.1</td>
<td>95.1</td>
</tr>
</tbody>
</table>

Fig. 7: ImageNet Architectures from Trade-Off Front.

Fig. 8: MAdds vs. ImageNet Accuracy. NATNets outperform other models in both objectives. In particular, NAT-M4 achieves a new state-of-the-art top-1 accuracy of 80.5% under mobile setting (≤600M MAdds). NAT-M1 improves MobileNetV3 top-1 accuracy by 2.3% with similar #MAdds.

### 4.4 Scalability to Datasets

Existing NAS approaches are rarely applied to datasets beyond standard ones (i.e. CIFAR-10 [9] and ImageNet [1]), where the classification task is at superordinate-level and the # of training images are sufficiently large. Instead, they adopt a conventional transfer learning setup [7], in which the architectures found by searching on standard benchmark datasets are transferred as is, with weights fine-tuned to new datasets. We argue that such a process is conceptually contradictory to the goal of NAS. The architectures transferred from standard datasets are sub-optimal either with respect to accuracy, efficiency or both. On the other hand, by transferring both architecture and weights NAT can indeed design bespoke models for each dataset.

We evaluated NAT on ten image classification datasets (see Table [5]) that present different challenges in terms of diversity in classes (superordinate vs. fine-grained) and size of training set (large vs small). For each dataset, we run NAT with two

#MAdds-accuracy trade-off curve comparison between NAT and existing NAS methods.

Results indicate that NATNets completely dominate (i.e. better in both #MAdds and accuracy) all existing designs, both manual and from other NAS algorithms, under mobile settings (≤600M MAdds). Compared to manually, designed networks, NAT is noticeably more efficient. NAT-M1 is 2.3% and 1.5% more accurate than MobileNetV3 [22] and FBNetV2-F4 [32] respectively, while being equivalent in efficiency (i.e. #MAdds, CPU and GPU latency). Furthermore, NATNets are consistently 6% more accurate than MobileNetV2 [56] scaled by width multiplier from 200M to 600M #MAdds. Our largest model, NAT-M4, achieves a new state-of-the-art ImageNet top-1 accuracy of 80.5% under mobile settings (≤600M MAdds). Interestingly, even though this experiment did not explicitly optimize for CPU or GPU latency, NATNets are faster than those (MobileNet-V3 [22], MnasNet [27]) that explicitly do optimize for latency.
Fig. 9: MAdds vs. Accuracy trade-off curves comparing NAT and existing architectures on a diverse set of datasets. The datasets are arranged in ascending order of training set size. Methods shown in the legend pre-train on ImageNet and fine-tune the weights on the target dataset. Methods with names annotated in sub-figures train from scratch or use external training data.

Fig. 10: Efficient architectures obtained by NAT (350M MAdds) on ten datasets.

objectives: maximize top-1 accuracy on validation data (20% randomly separated from the training set) and minimize #MAdds. We start from the supernet trained on ImageNet (which is created once before all experiments; see Section 4.2) and adapt it to the new dataset. During this procedure, the last linear layer is reset depending on the number of categories in the new dataset. NAT is now applied for a total of 30 iterations. In each iteration the supernet is adapted for 5 epochs using SGD with a momentum of 0.9. The learning rate is initialized at 0.01 and annealed to zero in 150 epochs (30 iterations with five epochs in each). All hyperparameters are set at default values from Table 2. For each dataset, the overall NAT process takes slightly under a day on a server with eight 2080Ti GPUs.

Fig. 9 shows the accuracy and #MAdds trade-off for each dataset across a wide range of models, including NATNets, existing NAS and hand-designed models. Across all datasets, NATNets consistently achieve better accuracy while being an order of magnitude more efficient (#MAdds) than existing models, suggesting that searching directly on the targeted datasets is a more effective alternative to the conventional transfer learning that fine-tunes weights of architectures learned on standard datasets (i.e. ImageNet and CIFAR-10). Under mobile settings ($\leq 600M$), NATNets achieve the state-of-the-art on these datasets, and a new state-of-the-art accuracy on both STL-10 [13] and CINIC-10 [12] datasets. Surprisingly, on small scale datasets e.g. Oxford Flowers102 [19], Oxford-IIT Pets [18], DTD [17] and STL-10 [13], we observe that NATNets are significantly more effective than conventional fine-tuning. Even on fine-grained datasets such as Stanford Cars and FGVC aircraft, where conventional fine-tuning did not improve upon training from scratch, NATNets improve accuracy while also being significantly more efficient.

Fig. 10 shows a visualization of architectures with 350M MAdds for each dataset. The lack of similarity in the networks suggest that different datasets require different architectures to be efficient in accuracy-MAdds, and NAT is able to generate these customized networks for each dataset. Additional visualization of

7. According to [70] for STL-10, and [71] for CINIC-10.
For better visualization we only show the architectures that are row. The 2D projection is obtained by ignoring the third objective. Each one of the four considered efficiency metrics in the bottom objective (such, as a weighted-sum of objectives [26], [27]). This is one of the advantages of using a many-objective optimization algorithm compared to optimizing a single scalarized objective.

More efficient in #MAdds or latency than another model with more #params and model efficiency (#MAdds and latency). In other words, trade-offs exist between model size (#params) and model efficiency (#MAdds, CPU latency, GPU latency). In total, we run three instances of three-objective metrics to quantify computational efficiency—#MAdds, CPU latency, GPU latency. We follow the settings from the ImageNet experiment in Section 4.3, except the fine-tuning step.

After obtaining the non-dominated (trade-off) solutions, we first visualize the objectives in Fig. 11. We observe that Pareto surfaces emerge at higher model complexity regime (i.e. top right corner) suggesting that trade-offs exist between model size (#params) and model efficiency (#MAdds and latency). We use NAT to simultaneously optimize for three objectives—namely, model accuracy on ImageNet, model size (#params), and model computational efficiency. We consider three different metrics to quantify computational efficiency—#MAdds, CPU latency, and GPU latency. In total, we run three instances of three-objective search—i.e. maximize accuracy, minimize #params, and minimize one of #MAdds, CPU latency or GPU latency. We follow the settings from the ImageNet experiment in Section 4.3 except the fine-tuning step.

Fig. 11 visualizes, in 2D, the top-1 accuracy as a trade-off with each one of the four considered efficiency metrics in the bottom row. The 2D projection is obtained by ignoring the third objective. For better visualization we only show the architectures that are close to the performance trade-off of MobileNetV3 [22], NATNets obtained directly from the three-objective search i.e., before any fine-tuning of their weights, consistently outperform MobileNetV3 on ImageNet along all the objectives (top-1 accuracy, #params, #MAdds, CPU and GPU latency). Additionally, we compare to MUXNets [54] which are also obtained from a three-objective NAS optimizing {top-1 accuracy, #params, and #MAdds}. However, MUXNets adopt a search space that is specifically tailored for reducing model size. Therefore, in comparison to MUXNets, we observe that NATNets perform favourably on all the remaining three efficiency metrics, except for #params. Primarily driven by curiosity in terms of pushing the scalability of our approach with respect to number of objectives, we provide an application to 12 objective problem in the supplementary materials under Section 6.

**4.5 Scalability to Objectives**

Practical applications of NAS can rarely be considered from the point of view of a single objective, and most often, they must be evaluated from many different, possibly competing, objectives. We demonstrate the scalability of NAT to more than two objectives, and evaluate its effectiveness.

We use NAT to simultaneously optimize for three objectives—namely, model accuracy on ImageNet, model size (#params), and model computational efficiency. We consider three different metrics to quantify computational efficiency—#MAdds, CPU latency, and GPU latency. In total, we run three instances of three-objective search—i.e. maximize accuracy, minimize #params, and minimize one of #MAdds, CPU latency or GPU latency. We follow the settings from the ImageNet experiment in Section 4.3 except the fine-tuning step.

Fig. 11 visualizes, in 2D, the top-1 accuracy as a trade-off with each one of the four considered efficiency metrics in the bottom row. The 2D projection is obtained by ignoring the third objective. For better visualization we only show the architectures that are close to the performance trade-off of MobileNetV3 [22], NATNets obtained directly from the three-objective search i.e., before any fine-tuning of their weights, consistently outperform MobileNetV3 on ImageNet along all the objectives (top-1 accuracy, #params, #MAdds, CPU and GPU latency). Additionally, we compare to MUXNets [54] which are also obtained from a three-objective NAS optimizing {top-1 accuracy, #params, and #MAdds}. However, MUXNets adopt a search space that is specifically tailored for reducing model size. Therefore, in comparison to MUXNets, we observe that NATNets perform favourably on all the remaining three efficiency metrics, except for #params. Primarily driven by curiosity in terms of pushing the scalability of our approach with respect to number of objectives, we provide an application to 12 objective problem in the supplementary materials under Section 6.

**4.6 Utility on Dense Image Prediction**

Dense image prediction is another series of important computer vision tasks, that assigns a label to each pixel in the input image [72], [73]. Success in these tasks relies on both feature extraction via a backbone CNN, e.g. ResNet [3], and feature aggregation, e.g. FPN [74], at multiple scales. In this section, we use NAT to design efficient backbone feature extractors for semantic segmentation, to demonstrate its utility beyond image classification.

Similar to previous studies, we start from the supernet trained on ImageNet (which is created once before all experiments; see Section 4.2). We remove the last classification layer and pair it with the BiSeNet segmentation heads [75], a lightweight semantic segmentation framework for real-time performance. We modify the searched input resolutions from [192, . . . , 256] to [512, . . . , 1280] and keep other searched options the same as before. NAT is applied to minimize #MAdds and maximize mIoU on validation data (20% randomly sampled from the training set) for 20 iterations. In each iteration, the supernet is adapted for 2K iterations using SGD with a momentum of 0.9 and weight

---

**Fig. 11:** Top row: NATNets obtained from tri-objective search to maximize ImageNet top-1 accuracy, minimize model size (#Params), and minimize {#MAdds, CPU latency, GPU latency} from left to right. Pareto surfaces emerge at higher model complexity regime (i.e. top right corner) suggesting that trade-offs exist between model size (#params) and model efficiency (#MAdds and latency). Bottom row: 2D projections from above 3D scatter, showing top-1 accuracy vs. each of the four efficiency related measurements. The first two 2D projections are from the first 3D scatter, and the remaining two 2D projections are from the second and third 3D scatters, respectively. To better visualize (the comparison with MobileNetV3 [22] and MUXNet [54]), partial solutions from the non-dominated frontiers are shown. All top-1 accuracy shown are without fine-tuning.
decay of $5 \times 10^{-4}$. We use a batch size of eight for each GPU. We use an initial learning rate of 0.01 and follow the “poly” learning rate schedule from the original BiSeNet [75], in which the initial learning rate is multiplied by $(1 - \frac{step}{max\_step})^{0.9}$ in each iteration. All other hyperparameters are set at default values from Table 2. On the Cityscapes dataset [20], the overall NAT process takes a day on a server with six Titan RTX GPUs. To realize the full potential of the searched NATNets, we further fine-tune the obtained models for 4K iterations. As shown in Table 5, the resulting NAT model yields comparable performance against state-of-the-art methods, including PSPNet [76], DeepLabv3 [77], Auto-DeepLab-S [21], while being 4x - 28x more efficient in #Madds.

**TABLE 5: Cityscapes Semantic Segmentation [20]:** All results are based on single-scale inputs from validation set.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Params</th>
<th>#Multi-Adds</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiSeNet [75]</td>
<td>13.4M</td>
<td>67B</td>
<td>74.8</td>
</tr>
<tr>
<td>PSPNet [76]</td>
<td>65.9M</td>
<td>2,017B</td>
<td>78.4</td>
</tr>
<tr>
<td>DeepLabV3+ [77]</td>
<td>43.5M</td>
<td>1,551B</td>
<td>79.6</td>
</tr>
<tr>
<td>Auto-DeepLab-S [21]</td>
<td>10.2M</td>
<td>333B</td>
<td>79.7</td>
</tr>
<tr>
<td>NAT + BiSeNet (ours)</td>
<td><strong>8.8M</strong></td>
<td><strong>73B</strong></td>
<td><strong>79.7</strong></td>
</tr>
</tbody>
</table>

## 5 ABLATION STUDY

In this section, we provide additional experiments towards quantifying the impacts of the main components introduced in NAT and hyperparameter analysis.

### 5.1 Accuracy Predictor Performance

In this subsection, we assess the effectiveness of different accuracy predictor models. We first uniformly sampled 350 architectures from our search space and trained them using SGD for 150 epochs on ImageNet. Each one of them is fine-tuned for 50 epochs on the other ten datasets (Table 3). From the 350 pairs of architectures and top-1 accuracy computed on each dataset, we reserved a subset (randomly chosen) of 50 pairs for testing, and the remaining 300 pairs are then available for training the predictor models.

Fig. 12 compares the mean (over 11 datasets) Spearman rank correlation between the predicted and the true accuracy for each accuracy predictor as the training sample size is varied from 50 to 300. Empirically, we observe that radial basis function (RBF) has higher Spearman rank correlation compared to the other three models. The proposed RBF ensemble model further improves performance over the standalone RBF model across all training sample size regimes. Fig. 12 shows a visualization of the comparative performance of predictor models on different datasets. From the trade-off perspective of minimizing number of training examples (which reduces the overall computational cost) and maximizing Spearman rank correlation in prediction (which improves the accuracy in ranking architectures during search), we chose the RBF ensemble as our accuracy predictor model and a training size of 100 for all our experiments.

### 5.2 Search Efficiency

The overall computation cost consumed by a NAS algorithm can be factored into three phases: (1) **Prior-search:** Cost incurred prior to architecture search, e.g., training supernet in case of one-shot approaches [10], [31] or constructing accuracy predictor [29], etc; (2) **During-search:** Cost associated with measuring the performance of candidate architectures sampled during search through inference. It also includes the cost of training the supernet in case it is coupled with the search, like in [24] and NAT; (3) **Post-search:** Cost associated with choosing a final architecture,
and/or fine-tuning/re-training the final architectures from scratch. For comparison, we select representative NAS algorithms, including those based on reinforcement learning (RL), gradients, evolutionary algorithm (EA), and weight sharing (WS). Table 6 shows results for ImageNet and CIFAR-10. The former is the dataset on which the supernet is trained and the latter is a proxy for transfer learning to a non-standard dataset. NAT consistently achieves better performance, both in terms of top-1 accuracy and model efficiency (e.g. #MAdds), compared to the baselines while computational cost is similar or lower. The primary computational cost of NAT is the prior-search training of supernet for 1200 hours. We emphasize, again, that it is a one-time cost that is amortized across all subsequent deployment scenarios (e.g. 10 additional datasets in Section 4.4).

### 5.3 Analysis of Crossover

Crossover is a standard operator in evolutionary algorithms, but has largely been avoided by existing EA-based NAS methods [27], [38], [39]. But as we demonstrate here, a carefully designed crossover operation can significantly improve search efficiency. We run the evolutionary search of NAT with and without the crossover operator on four datasets; ImageNet [1], CIFAR-10 [9], CIFAR-100, and Stanford Cars [15]. The hyperparameters that we compare are:

1. **w/ crx**: crossover probability of 0.9; mutation probability of 0.1; mutation index $\eta_m$ of 3.
2. **w/o crx**: crossover probability of 0.0; mutation probability of 0.2; mutation index $\eta_m$ of 3.

We double the mutation probability when crossover is not used to compensate for the reduced exploration ability of the search. On each dataset, we run each setting to maximize the top-1 accuracy 11 times and report the median performance as a function of the number of architectures sampled in Fig. 14a. On all four datasets, the crossover operator significantly improves the efficiency of the evolutionary search algorithm. To further validate, we sweep over the probability of crossover while maintaining the rest of the settings. The median performance (over 11 runs) deteriorates as the crossover probability is reduced from 0.9 to 0.2 (see Fig. 14b).

### 5.4 Analysis of Mutation Hyperparameters

The mutation operator used in NAT is controlled via two hyperparameters—namely, the mutation probability $p_m$ and mutation index $\eta_m$. To identify the optimal hyperparameter values, we conduct the following parameter sweep experiments. Setting the rest of the hyperparameters to their default values (see Table 2), we sweep the value of $p_m$ from 0.1 to 0.8 and $\eta_m$ from 1.0 to 20. And for each setting, we run NAT eleven times on four datasets (same as the crossover experiment) to maximize the top-1 accuracy. Figs. 15a and 15b show the effect of mutation probability $p_m$ and mutation index $\eta_m$, respectively. We observe that increasing the mutation probability has an adverse effect on performance. Similarly, low values of $\eta_m$, which encourages the mutated offsprings to be further away from parent architectures, improves the performance. Based on these observations, we set the mutation probability $p_m$ and mutation index $\eta_m$ parameters to 0.1 and 1.0, respectively, for all our experiments in Section 4.

### 5.5 Effectiveness of Supernet Adaptation

Recall that NAT adopts any supernet trained on a large-scale dataset, e.g. ImageNet, and seeks to efficiently transfer to a task-specific supernet on a given dataset. Here, we compare this procedure to a more conventional approach of adapting every subnet (candidate architectures in search) directly. Specifically, we consider the following.

1. **Supernet Adaptation**: fine-tune supernet for 5 epochs in each iteration and use accuracy from inherited weights (without further training) to select architectures during search (adopted in NAT).
2. **Subset Adaptation**: fine-tune each subnet for 5 epochs from the inherited weights, then measure the accuracy.

We apply these two approaches to a bi-objective search of maximizing top-1 accuracy and minimizing #MAdds on four datasets, including CIFAR-10, CIFAR-100, Oxford Flowers102,
and STL-10. Figure 16 compares the final Pareto fronts. Adapting the supernet yields significantly better performance than adapting individual subnets. Furthermore, we select a subset of searched subnets after subnet adaptation and fine-tune their weights for an additional 150 epochs. We refer to this as additional fine-tuning in Fig. 16. Empirically, we observe that further fine-tuning can match the performance of supernet adaptation on datasets with larger training samples per class (e.g., 4,000 in CIFAR-10). On datasets with fewer samples per class (e.g., 20 in Flowers 102), there is still a large performance gap between supernet adaptation and additional fine-tuning. Overall, the results suggest that supernet adaptation is more effective on tasks with limited training samples.

5.6 Towards Quantifying Architectural Advancement

Comparing the architectural contribution to the success of different NAS algorithms can be difficult and ambiguous due to substantial differences in training procedures, e.g., data augmentation, training hyperparameters, etc. Therefore, to quantify the architectural advancement made by NAT alone, we train NAT-M1 from randomly initialized weights (instead of inheriting them from the supernet) with standard training hyperparameters (see Table 7). We then compare the outcome to two other recently proposed efficient models, MobileNetV3 [22] and FBNNetV2 [32]. The results are summarized in Table 8 where we observe that the NAT searched model, NAT-M1, is 0.5 - 1.0% more accurate on ImageNet than compared models using similar or less #MAdds.

**TABLE 7:** Details of training hyperparameter settings. Advance settings are in addition to standard settings.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Data Augmentation</th>
<th>Regularization</th>
<th>Optimizer</th>
<th>LR Schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>Horizontal Flip + Crop</td>
<td>Drop out</td>
<td>RMSProp + Exponential Moving Average</td>
<td>Step LR w/ Decay + Linear Warm-up</td>
</tr>
<tr>
<td>Advance</td>
<td>+ Random Augmentation</td>
<td>+ Random Erase Pixel</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To further quantify the architectural advancement made by NAT, we use NAT-M1 as a drop-in replacement of the backbone feature extractor for three dense image prediction tasks, including object detection, semantic segmentation, and instance segmentation. More specifically, we replace the EfficientNet-B0 [28] in EfficientDet-D0 [84] for object detection; the ResNet-18 [3] in BiSeNet [75] for semantic segmentation; and the ResNet-50 [3] in YOLACT [85] for instance segmentation. For comparison, we apply the same procedure to both MobileNetV3 and FBNNetV2.
Finally, we break down the effect of different training settings and additional fine-tuning for the Top-1 accuracy of the searched models in Table 9. The advance setting in Table 7 also uses knowledge distillation [31], [68].

TABLE 9: Effect of different training setups. Details of the standard and advanced settings under Random Initialization are provided in Table 7.

<table>
<thead>
<tr>
<th>Training Settings</th>
<th>Random Initialization</th>
<th>Inherited from SuperNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>standard</td>
<td>advanced</td>
</tr>
<tr>
<td>NAT-M1</td>
<td>75.7</td>
<td>77.1</td>
</tr>
<tr>
<td>NAT-M2</td>
<td>76.9</td>
<td>78.0</td>
</tr>
<tr>
<td>NAT-M3</td>
<td>78.2</td>
<td>79.1</td>
</tr>
<tr>
<td>NAT-M4</td>
<td>78.8</td>
<td>79.5</td>
</tr>
</tbody>
</table>

6 Conclusion

This paper considered the problem of designing custom neural network architectures that trade-off multiple objectives for a given image classification task. We introduced Neural Architecture Transfer (NAT), a practical and effective approach for this purpose. We described our efforts to harness the concept of a supernet and an evolutionary search algorithm for designing task-specific neural networks trading-off accuracy and computational complexity. We also showed how to use an online regressor, as a surrogate model to predict the accuracy of subnets in the supernet. Experimental evaluation on eleven benchmark image classification datasets, ranging from large-scale multi-class to small-scale fine-grained tasks, showed that networks obtained by NAT outperform conventional fine-tuning based transfer learning, while being orders of magnitude more efficient under mobile settings (≤ 600M Multiply-Adds). NAT was especially effective for small-scale fine-grained tasks where fine-tuning pre-trained ImageNet models is ineffective. Finally, we also demonstrated the utility of NAT in optimizing up to twelve objectives with a subsequent trade-off analysis procedure for identifying a single preferred solution. Overall, NAT is the first large scale demonstration of many-objective neural architecture search for designing custom task-specific models on diverse image classification datasets.

REFERENCES

Zhichao Lu received his B.Sc and Ph.D. degrees in Electrical and Computer Engineering from Michigan State University, East Lansing, USA, in 2013 and 2020. He is currently a postdoc research fellow with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. His research interests are in the field of evolutionary machine learning assisted evolutionary algorithms, automated machine learning, and in particular evolutionary neural architecture search. He received the best paper award at GECCO 2019 under evolutionary machine learning track.

Gautam Sreekumar is a second year Ph.D. student in the Department of Computer Science and Engineering in Michigan State University. Previously, he obtained his B.Tech. degree in Electrical Engineering from Indian Institute of Technology Madras. His research interests are in the field of deep neural networks, particularly evolutionary neural architecture search for vision tasks.

Erik D. Goodman is PI and Executive Director of the BEACON Center for the Study of Evolution in Action, an NSF Science and Technology Center headquartered at Michigan State University, funded at $47.5 million for 2010-20. Ph.D., Computer and Communication Sciences, University of Michigan, 1972. Asst. Prof.-Full Prof., Michigan State University, 1971-. Electrical and Computer Engineering, Mechanical Engineering, Computer Science and Engineering. Director, Case Center for Computer-Aided Engineering and Manufacturing, 1983-2002; Director. Co-founder and VP Technology, Red Cedar Technology, Inc., which develops design optimization software, now owned by Siemens. Chair and Senior Fellow, International Society for Genetic and Evolutionary Computation; Founding Chair, ACM SIG on Genetic and Evolutionary Computation (SIGEVO), 2005.

Kalyanmoy Deb is the Koenig Endowed Chair Professor with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan, USA. He received his Bachelor’s degree in Mechanical Engineering from IIT Kharagpur in India, and his Master’s and Ph.D. degrees from the University of Alabama, Tuscaloosa, USA, in 1989 and 1991. He is largely known for his seminal research in evolutionary multi-criterion optimization. He has published over 544 international journal and conference research papers. His current research interests include evolutionary optimization and its application in design, modeling, AI, and machine learning. Dr. Deb is a recipient of the IEEE CIS EC Pioneer Award in 2018, the Lifetime Achievement Award from Clarivate Analytics in 2017, the Infosys Prize in 2012, and the Edgeworth-Pareto Award in 2008. He is a fellow of IEEE and ASME. More information about his work can be found from http://www.coin-lab.org.

Vishnu Naresh Boddeti is an Assistant Professor in the computer science department at Michigan State University. He received a Ph.D. degree in Electrical and Computer Engineering program at Carnegie Mellon University in 2013. His research interests are in Computer Vision, Pattern Recognition and Machine Learning. He received the best paper award at BTAS 2013, the best student paper award at ACCV 2018, and the best paper award at GECCO 2019.
Neural Architecture Transfer
(Supplementary Material)

Zhichao Lu, Student Member, IEEE, Gautam Sreekumar, Erik Goodman, Wolfgang Banzhaf, Kalyanmoy Deb, Fellow, IEEE and Vishnu Naresh Boddeti, Member, IEEE

In this supplementary material, we include: (i) an in-depth discussion relating our proposed approach to existing one-shot NAS approaches in Section 1; (ii) a comparison between domination and reference point based many-objective selection in Section 2; (iii) an “unsupervised” method for choosing an architecture based on trade-offs in Section 3; (iv) additional comparison to existing convolutional neural networks on ImageNet [1] in Section 4; (v) a visualization of the obtained architectures in Section 5; and (vi) a twelve-objective use case of NAT.

1 Connection to Existing One-Shot NAS

Most existing one-shot NAS approaches follow a two-step process, where the supernet training and the architecture search are disentangled into two sequential stages. This process starts with training a supernet (in which searchable architectures become subnets) offline as a one-time process prior to the search. Then the performance of the subnets, evaluated with the inherited weights, is used to guide the selection of architectures during search. Early one-shot approaches [2], [3], [4] follow a conventional (rather naïve) way to train the supernet, i.e. train a randomly chosen subpart (subnet) of the supernet directly from randomly initialized weights for each mini-batch (see Fig. 4a). Consequently, the searched subnets need to be re-trained thoroughly from scratch as the performance evaluated with inherited weights are far below the true performance and can only be used as a proxy indicator to compare the relative difference between subnets.

The progressive shrinking algorithm proposed in OnceForAll [5] also trains the supernet in an offline fashion, but differs in three aspects—(i) it pre-trains the supernet at full scale before sampling subnets; (ii) it gradually adds the searched dimensions (kernel size, depth, width) into the search space; and (iii) it uses the full-scale supernet to supervise the training of subnets. However, the supernet weights update is still based on randomly sampled subnets. See Fig. 4a for a visualization. Empirically, OnceForAll shows that the supernet trained with progressive shrinking enables subnets with inherited weights to be directly deployed without re-training.

Despite the success shown in OnceForAll, we argue that such an offline training process of supernet is fundamentally limited by the fact that it requires all subnets to be learned simultaneously. To elaborate, without prior knowledge on the distribution of the optimal subnets for the tasks at hand, the supernet training has to cover the search space of subnets globally as the training is performed prior to the search as a one-time process. However, training the supernet weights in such a way that all subnets are optimized simultaneously is practically infeasible. For instance, progressive shrinking [5] sampled roughly 634K subnets during supernet training, which is less than $10^{-12}\%$ of the its total subnet volume. Any additional options added to the search space (one more kernel size and expand ratio choice) will require 100x more training epochs (100K vs 1K) to cover the same volume of subnets, which is obviously not scalable. Moreover, we argue that simultaneously training all subnets is also unnecessary as not all subnets are equally important for the tasks at hand. Specifically, existing NAS works have shown that different hardware requires different architectures to be efficient, e.g. CPU favors deeper networks with fewer channels in each layer, while GPU favors shallower networks with more channels in each layer, from the latency perspective [6], [7].

To overcome the aforementioned limitations of existing one-shot approaches, we propose NAT. The key difference is that NAT trains the supernet online. Instead of randomly sampling subnets to train the supernet all at once, NAT estimates the distribution (in the variable space) of the optimal subnets from the subnets returned by a many-objective search algorithm, and trains the supernet in correspondence to the estimated distribution. NAT does so in a progressive manner, where the estimated distribution and supernet training are gradually refined through iterations (see Fig. 2). We argue that our approach is conceptually more scalable and efficient than existing one-shot approaches since the supernet training now can focus on the promising task-specific subnets recommended by the search algorithm, instead of on all subnets globally.

To visualize the difference between the existing approach of disentangling supernet training from architecture search, and our approach that use architecture search to guide the supernet training, let us consider the following problem of minimizing a two-variable Rosenbrock function [8]:

$$\text{minimize } f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2,$$
$$x_1, x_2 \in [-2.048, 2.048].$$

The objective landscape (contour) of the above two-variable Rosenbrock function is shown in Fig. 5a. Let’s also assume that each function evaluation of $f(x_1, x_2)$ in Eq (1) is expensive and hence extensively probing the true value is prohibitive (as in the case of NAS). To efficiently optimize this problem, we may learn a

1. Estimated based on the batch size of 2,048 and the training epochs of 1,000 provided by the OnceForAll paper [5].
Recall from Section 3.5 in the main paper that domination resides, eventually leading to a better solution. Local regions where potential optimal solutions are more likely to appear are lexicographically better than solutions in higher rank; and solutions in the same rank are non-dominated, i.e. equally good. To compensate for the degradation in selection pressure from domination alone, many recently proposed many-objective optimization algorithms \cite{9}, \cite{10}, \cite{11}, \cite{12} opt for the route of reference point based selection, including this work. The reference points serve as a set of pre-defined targets to aid the selection. See Fig. 5 for a visualization. As shown in Fig. 4, the online approach allows the meta-model to focus on local regions where potential optimal solutions are more likely to reside, eventually leading to a better solution.

2 Many-Objective Selection Continued

Recall from Section 3.5 in the main paper that domination is a widely-adopted partial ordering concept to compare solutions with two or more objectives. It is used to sort solutions into different ranks of importance, where solutions in lower rank are lexicographically better than solutions in higher rank; and solutions in the same rank are non-dominated, i.e. equally good. However, as well recognized by the evolutionary many-objective optimization community \cite{9}, \cite{10}, an increasing larger fraction of randomly generated solutions becomes non-dominated as the number of objectives increases (see Fig. 5 for a visualization). As a result, the selection pressure provided from domination diminishes quickly as the number of objectives increases, leading to a slow convergence towards the Pareto front.

To compensate for the degradation in selection pressure from domination alone, many recently proposed many-objective optimization algorithms \cite{9}, \cite{10}, \cite{11}, \cite{12} opt for the route of reference point based selection, including this work. The reference points serve as a set of pre-defined targets to aid the selection. See Fig. 5 for a visualization. As shown in Fig. 4, the online approach allows the meta-model to focus on local regions where potential optimal solutions are more likely to reside, eventually leading to a better solution.

3 Choosing Best Trade-off Solution

The proposed many-objective EA is expected to produce $N$ (population size) solutions trading-off all $m$ objectives. The performance assessment indicator for comparing many-objective optimization algorithms. We vary the number of objectives in DTLZ1 from 3 to 15 and perform 31 independent runs for each selection method. The mean IGD values along with the standard deviations are plotted in Fig. 6. The consistently lower IGD values across different numbers of objectives confirm the effectiveness of the reference point based selection method.

2 Note that Hypervolume, another multi-objective performance metric that is used in the main paper, is computationally infeasible to calculate under large numbers of objectives \cite{13}. 

Fig. 2: Overview of our proposed NAT. The distribution of optimal subnets is estimated from the promising architectures returned by the architecture search. Then it is used to guide the training of the supernet. The “per iteration” refers to the iteration in Algorithm 1 in the main paper.

Fig. 1: Overview of existing one-shot NAS approaches, which decouples the supernet training and architecture search to two sequential steps.
Evaluated solutions ★ Promising solution to be evaluated ★ Optimum obtained ★ True optimum

Fig. 3: (a) True objective landscape (contour) of a two-variable Rosenbrock function. (b) Offline surrogate modelling approach (adopted by existing one-shot NAS methods [3], [4], [5]): the objective landscape is interpolated through uniformly sampled solutions, then the optimization is carried out on the interpolated landscape. (c) Online surrogate modelling approach (ours): a coarse interpolation of the objective landscape is firstly learned using partial budget, then the landscape is gradually refined by adding the optimization outcome on the current landscape to the interpolation. See Fig. 4 for comparison on the obtained results.

Fig. 4: Top row compares the interpolated landscapes and the obtained optimum by (a) our online surrogate modeling (Fig. 3c) with initial, 3/4, and full budget from Left to Right, and (b) offline surrogate modeling (existing one-shot NAS approaches; Fig. 3b). Bottom row visualizes the evaluated solutions by the two approaches. Even though the offline approach of uniformly sampling provides a better global interpolation of the landscape (i.e. sub-figure (b)), the online approach achieves a better local interpolation around the optimum (i.e. sub-figure (a) Right). The true landscape is shown in Fig. 3a.

Thereafter, the solutions having the highest trade-off value indicates that it causes the largest average loss in some objectives to make a unit average gain in other objectives to choose any of its neighbors. If this highest trade-off value is much larger statistically than other solutions, then the highest trade-off solution is the preferred choice, in case of no preferences provided from users.

solutions are guaranteed to have one property: a gain in one objective between \(i\)-th and \(j\)-th solutions comes only from a loss in at least one other objective between them. We calculate the trade-off of \(i\)-th solution as the average loss per unit average gain among \(m\) nearest neighbors \((B(i))\) based on normalized Euclidean distance are used here), as follows [17]:

\[
\text{Trade-off}(i) = \frac{B(i)}{\max_{j=1}^{M} \frac{\text{Avg.Loss}(i,j)}{\text{Avg.Gain}(i,j)}}
\]

where

\[
\text{Avg.Loss}(i,j) = \frac{\sum_{k=1}^{M} \max(0, f_k(j) - f_k(i))}{\sum_{k=1}^{M} \{1|f_k(j) > f_k(i)|\}}
\]

\[
\text{Avg.Gain}(i,j) = \frac{\sum_{k=1}^{M} \max(0, f_k(i) - f_k(j))}{\sum_{k=1}^{M} \{1|f_k(i) > f_k(j)|\}}
\]
TABLE 1: NAT model performance corresponding to Fig. 9 in main paper.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Params</th>
<th>#MAdds</th>
<th>Top-1 Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxford-IIIT Pets</td>
<td>3.3M</td>
<td>125M</td>
<td>97.5</td>
</tr>
<tr>
<td></td>
<td>3.4M</td>
<td>195M</td>
<td>97.9</td>
</tr>
<tr>
<td></td>
<td>3.7M</td>
<td>250M</td>
<td>98.1</td>
</tr>
<tr>
<td></td>
<td>4.2M</td>
<td>400M</td>
<td>98.3</td>
</tr>
<tr>
<td>DTD</td>
<td>4.0M</td>
<td>297M</td>
<td>77.6</td>
</tr>
<tr>
<td></td>
<td>4.3M</td>
<td>232M</td>
<td>97.4</td>
</tr>
<tr>
<td></td>
<td>4.1M</td>
<td>347M</td>
<td>78.4</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>3.8M</td>
<td>261M</td>
<td>86.0</td>
</tr>
<tr>
<td></td>
<td>6.9M</td>
<td>468M</td>
<td>98.4</td>
</tr>
<tr>
<td></td>
<td>4.1M</td>
<td>266M</td>
<td>88.5</td>
</tr>
<tr>
<td>STL10</td>
<td>5.1M</td>
<td>388M</td>
<td>90.1</td>
</tr>
<tr>
<td></td>
<td>6.4M</td>
<td>398M</td>
<td>87.5</td>
</tr>
<tr>
<td></td>
<td>2.2M</td>
<td>136M</td>
<td>76.1</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>4.6M</td>
<td>291M</td>
<td>97.9</td>
</tr>
<tr>
<td></td>
<td>3.1M</td>
<td>198M</td>
<td>87.4</td>
</tr>
<tr>
<td></td>
<td>7.5M</td>
<td>573M</td>
<td>97.9</td>
</tr>
<tr>
<td></td>
<td>6.2M</td>
<td>411M</td>
<td>94.1</td>
</tr>
<tr>
<td></td>
<td>3.9M</td>
<td>299M</td>
<td>89.0</td>
</tr>
<tr>
<td>CINIC-10</td>
<td>9.1M</td>
<td>710M</td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>4.4M</td>
<td>240M</td>
<td>96.7</td>
</tr>
<tr>
<td></td>
<td>5.1M</td>
<td>303M</td>
<td>97.2</td>
</tr>
<tr>
<td></td>
<td>7.5M</td>
<td>436M</td>
<td>97.8</td>
</tr>
<tr>
<td></td>
<td>5.7M</td>
<td>471M</td>
<td>94.1</td>
</tr>
<tr>
<td></td>
<td>3.2M</td>
<td>175M</td>
<td>87.0</td>
</tr>
<tr>
<td></td>
<td>4.6M</td>
<td>317M</td>
<td>93.4</td>
</tr>
<tr>
<td>Flowers102</td>
<td>7.8M</td>
<td>492M</td>
<td>87.7</td>
</tr>
<tr>
<td></td>
<td>6.3M</td>
<td>560M</td>
<td>79.1</td>
</tr>
<tr>
<td></td>
<td>4.0M</td>
<td>160M</td>
<td>91.8</td>
</tr>
<tr>
<td></td>
<td>9.0M</td>
<td>796M</td>
<td>88.3</td>
</tr>
<tr>
<td></td>
<td>4.2M</td>
<td>240M</td>
<td>96.7</td>
</tr>
<tr>
<td></td>
<td>3.7M</td>
<td>250M</td>
<td>98.1</td>
</tr>
<tr>
<td></td>
<td>5.3M</td>
<td>474M</td>
<td>97.4</td>
</tr>
<tr>
<td></td>
<td>5.0M</td>
<td>315M</td>
<td>96.7</td>
</tr>
<tr>
<td></td>
<td>4.6M</td>
<td>317M</td>
<td>93.4</td>
</tr>
<tr>
<td>FGVC Aircraft</td>
<td>6.2M</td>
<td>392M</td>
<td>98.2</td>
</tr>
<tr>
<td></td>
<td>4.0M</td>
<td>297M</td>
<td>77.6</td>
</tr>
<tr>
<td></td>
<td>5.1M</td>
<td>303M</td>
<td>97.2</td>
</tr>
<tr>
<td></td>
<td>4.5M</td>
<td>200M</td>
<td>97.9</td>
</tr>
<tr>
<td></td>
<td>4.3M</td>
<td>232M</td>
<td>97.4</td>
</tr>
<tr>
<td></td>
<td>4.1M</td>
<td>347M</td>
<td>78.4</td>
</tr>
<tr>
<td></td>
<td>5.1M</td>
<td>303M</td>
<td>97.2</td>
</tr>
<tr>
<td></td>
<td>4.4M</td>
<td>240M</td>
<td>96.7</td>
</tr>
<tr>
<td></td>
<td>3.7M</td>
<td>240M</td>
<td>97.9</td>
</tr>
<tr>
<td></td>
<td>4.0M</td>
<td>297M</td>
<td>77.6</td>
</tr>
<tr>
<td></td>
<td>5.1M</td>
<td>303M</td>
<td>97.2</td>
</tr>
<tr>
<td></td>
<td>4.5M</td>
<td>232M</td>
<td>97.4</td>
</tr>
<tr>
<td></td>
<td>4.0M</td>
<td>297M</td>
<td>77.6</td>
</tr>
<tr>
<td></td>
<td>5.1M</td>
<td>303M</td>
<td>97.2</td>
</tr>
</tbody>
</table>

4 COMPARISON TO EXISTING CONVNETS

Figure 7 visualizes the #MAdds-accuracy trade-off curve, where our NATNets achieve better top-1 accuracy with much fewer #MAdds than other CNN models. Notably, NAT-M1 is more accurate, and 20x more efficient in #MAdds than ResNet-50 [18]; NAT-M4 is more accurate, and 21x more efficient in #MAdds than Inception-ResNet-v2 [19].

5 ARCHITECTURE VISUALIZATION

One of the main advantages of multi-objective optimization is that it generates a set of non-dominated solutions in a single run. These non-dominated solutions are special in the sense that one has to sacrifice on one objective to gain on another. Thereby, “mining” on these non-dominated solutions oftentimes yields important design principles for the task at hand, in this case, to efficiently construct an architecture specific to the objectives and dataset. To demonstrate this concept, we visualize the non-dominated architectures (to maximize top-1 accuracy and minimize #MAdds) resulting from NAT on a diverse set of datasets in Fig. 8. Each subfigure is a heat map showing the distribution of the searched, input image resolutions, width multipliers, and layer settings.

It is clear from Fig. 8 that even under the same objectives, the optimal architectures for different datasets are different. For example, the most frequent input image resolution is 192 (the lowest value in our searched options) for Oxford-IIIT Pets [21] and STL-10 [23]. While on FGVC Aircraft [24] and Food-101 [27], the most frequent choice of resolution is 256, which is the highest value in our searched option. Similar observations can be made in case of width multiplier and layer settings. This example provides empirical evidence necessary for finding dataset-specific optimal architectures, as opposed to conventional transfer learning. And as demonstrated in the main paper, our proposed NAT presents an efficient and effective way to achieve this goal.

6 SCALABILITY TO OBJECTIVES CONTINUED

To further validate the scalability of NAT to a large number of objectives, we consider the top-1 accuracy on each of the 11

Fig. 5: Mean ratio of non-dominated solutions from a set of randomly generated solutions. N is the sample size of the randomly generated solutions.

Fig. 6: Performance comparison of reference point based (Algorithm 4 in the main paper) and domination based selections [16] on DTLZ1 problem [13].

Fig. 7: #MAdds vs. ImageNet Accuracy. Our NATNets significantly outperform other models from NAS algorithms and human experts. In particular, NAT-M4 achieves new state-of-the-art 80.5% top-1 accuracy under mobile setting (600M MAdds).
The above analysis alludes to a computational mechanism for choosing a single preferred trade-off solution from the Pareto solutions obtained by a many-objective optimization algorithm. If such an overwhelmingly high trade-off solution exists in the Pareto front, it becomes one of the best choices and can outperform solutions found by a single-objective optimization algorithm. Without resorting to a many-objective optimization to find multiple trade-off solutions, identification of such a high trade-off solution is very challenging.

Fig. 8 (Left) shows the 12 objective values for all 45 non-dominated architectures obtained by NAT in a parallel coordinate plot (PCP), where each vertical bar is an objective and each line connecting all 12 vertical bars is an architecture. We now apply the trade-off decision analysis presented in the supplementary materials under Section 1 and observe that the highest trade-off solution is more than the trade-off away from the rest of 44 solutions. This solution is highlighted in dark blue in Fig. 9 (Left).

### Table 2: Accuracy predictor model mean (standard deviation) performance corresponding to Fig. 13 in main paper.

<table>
<thead>
<tr>
<th>Method</th>
<th>ImageNet</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
<th>Flowers102</th>
<th>Food-101</th>
<th>Oxford-IIT Pet</th>
<th>Aircraft</th>
<th>Stanford Cars</th>
<th>DTD</th>
<th>STL-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP</td>
<td>0.606 (0.09)</td>
<td>0.969 (0.01)</td>
<td>0.693 (0.13)</td>
<td>0.918 (0.02)</td>
<td>0.980 (0.04)</td>
<td>0.945 (0.02)</td>
<td>0.551 (0.17)</td>
<td>0.964 (0.01)</td>
<td>0.467 (0.11)</td>
<td>0.973 (0.11)</td>
</tr>
<tr>
<td>RBF</td>
<td>0.705 (0.11)</td>
<td>0.969 (0.01)</td>
<td>0.806 (0.08)</td>
<td>0.932 (0.03)</td>
<td>0.981 (0.01)</td>
<td>0.967 (0.01)</td>
<td>0.693 (0.08)</td>
<td>0.977 (0.01)</td>
<td>0.653 (0.06)</td>
<td>0.979 (0.01)</td>
</tr>
<tr>
<td>MLP</td>
<td>0.635 (0.09)</td>
<td>0.851 (0.06)</td>
<td>0.562 (0.10)</td>
<td>0.766 (0.06)</td>
<td>0.775 (0.09)</td>
<td>0.798 (0.05)</td>
<td>0.658 (0.15)</td>
<td>0.717 (0.10)</td>
<td>0.490 (0.09)</td>
<td>0.899 (0.06)</td>
</tr>
<tr>
<td>DT</td>
<td>0.625 (0.11)</td>
<td>0.974 (0.01)</td>
<td>0.736 (0.11)</td>
<td>0.940 (0.02)</td>
<td>0.990 (0.01)</td>
<td>0.961 (0.01)</td>
<td>0.629 (0.14)</td>
<td>0.986 (0.01)</td>
<td>0.590 (0.14)</td>
<td>0.976 (0.01)</td>
</tr>
<tr>
<td>RBF Ensemble</td>
<td>0.866 (0.04)</td>
<td>0.959 (0.02)</td>
<td>0.858 (0.05)</td>
<td>0.931 (0.01)</td>
<td>0.967 (0.03)</td>
<td>0.943 (0.01)</td>
<td>0.870 (0.07)</td>
<td>0.975 (0.01)</td>
<td>0.890 (0.04)</td>
<td>0.964 (0.02)</td>
</tr>
</tbody>
</table>

### References


Fig. 9: Left: Parallel Coordinate Plot (PCP) where each vertical bar is an objective and each line is a non-dominated architectures achieved by NAT from a 12-obj optimization of minimizing #MAdds and maximizing accuracy on the 11 datasets. The model with the best trade-off (see supplementary materials under Section 1 for details) is highlighted in dark blue. Right: 1-on-1 comparison between the selected NATNet (top-ranked in trade-off) and representative peer models on top-1 accuracy on various datasets. Method with larger area is better.


