BossNAS: Exploring Hybrid CNN-transformers with Block-wisely Self-supervised Neural Architecture Search

Changlin Li\textsuperscript{1}, Tao Tang\textsuperscript{2}, Guangrun Wang\textsuperscript{3,4}, Jiefeng Peng\textsuperscript{3}, Bing Wang\textsuperscript{5}, Xiaodan Liang\textsuperscript{2,3,}\textsuperscript{*}, Xiaojun Chang\textsuperscript{6}
\textsuperscript{1}GORSE Lab, Dept. of DSAI, Monash University \textsuperscript{2}Sun Yat-sen University \textsuperscript{3}DarkMatter AI Research \textsuperscript{4}University of Oxford \textsuperscript{5}Alibaba Group \textsuperscript{6}RMIT University
changlin.li@monash.edu, 
\{trent.tangtao, wanggrun, jiefengpeng, xdliang328\}@gmail.com, fengquan.wb@alibaba-inc.com, xiaojun.chang@rmit.edu.au

Abstract

A myriad of recent breakthroughs in hand-crafted neural architectures for visual recognition have highlighted the urgent need to explore hybrid architectures consisting of diversified building blocks. Meanwhile, neural architecture search methods are surging with an expectation to reduce human efforts. However, whether NAS methods can efficiently and effectively handle diversified search spaces with disparate candidates (e.g. CNNs and transformers) is still an open question. In this work, we present Block-wisely Self-supervised Neural Architecture Search (BossNAS), an unsupervised NAS method that addresses the problem of inaccurate architecture rating caused by large weight-sharing space and biased supervision in previous methods. More specifically, we factorize the search space into blocks and utilize a novel self-supervised training scheme, named ensemble bootstrapping, to train each block separately before searching them as a whole towards the population center. Additionally, we present HyTra search space, a fabric-like hybrid CNN-transformer search space with searchable downsampling positions. On this challenging search space, our searched model, BossNet-T, achieves up to 82.2% accuracy on ImageNet, surpassing EfficientNet by 2.1% with comparable compute time. Moreover, our method achieves superior architecture rating accuracy with 0.78 and 0.76 Spearman correlation on the canonical MBConv search space with ImageNet and on NATS-Bench size search space with CIFAR-100, respectively, surpassing state-of-the-art NAS methods.\textsuperscript{1}

1. Introduction

The development of neural network architectures has brought about significant progress in a wide range of visual recognition tasks over the past several years. Representative examples of such models include ResNet [24], SENet [30], MobileNet [29] and EfficientNet [63]. Recently, the newly emerging attention-based architectures are coming to the forefront in the vision field, challenging the dominance of convolutional neural networks (CNNs). This exciting breakthrough in vision transformers led by ViT [19] and DETR [8], are achieving competitive performance on various vision tasks, such as image classification [19, 65, 78, 13], object detection [8, 89, 61], semantic segmentation [87], and others [25, 50, 32]. As suggested by prior works [19, 61, 3], hybrids of CNNs and transformers can outperform both pure transformers and pure CNNs.

Despite the large advances brought about by network design, manually finding well-optimized hybrid architectures can be challenging, especially as the number of design choices increases. Neural Architecture Search (NAS) is a popular approach to reducing the human effort in network architecture design by automatically searching for optimal
architectures in a predefined search space. Representative success in performing NAS on manually designed building blocks include MobileNetV3 [28], EfficientNet [63], etc. These works are searched by multi-trial NAS methods [62, 91, 2, 88, 11, 47], which are computationally prohibitive (costing thousands of GPU days). Recent weight-sharing NAS methods [6, 52, 4, 43] encode the entire search space as a weight-sharing supernet to avoid repetitive training of candidate networks, thus largely reducing the search cost.

However, as shown in Fig. 1a, architecture search spaces with layer-level granularity grow exponentially with increased network depth, which has been identified (in [36, 39]) as the main culprit of inaccurate architecture rating in weight-sharing NAS methods. To reduce the size of the large weight-sharing space, previous works [36, 46] factorize the search space into blocks and use a pretrained teacher model to provide block-wise supervision (Fig. 1b). Despite their high ranking correlation and high efficiency, we find (in Sec. 5) their results to be highly correlated with the teacher architecture. As illustrated in Fig. 1b, when training by a teacher with blue nodes, candidate architectures with more blue nodes tend to get higher ranks in these methods. This limits its application on diversified search spaces with disparate candidates, such as CNNs and transformers.

On the other hand, unsupervised NAS [41] has recently emerged as an interesting research topic. Without access to any human-annotated labels, unsupervised NAS methods (optimized with pretext tasks [41] or random labels [86]) have been proven capable of achieving comparable performance to supervised NAS methods. Accordingly, we propose to use an unsupervised learning method as an alternative to supervised distillation in the aforementioned block-wise NAS scheme (Fig. 1c), aiming to address the problem of architectural bias caused by the use of the teacher model.

In this work, we propose a novel unsupervised NAS method, Block-wise Self-supervised Neural Architecture Search (BossNAS), which aims to address the problem of inaccurate predictive architecture ranking caused by a large weight-sharing space while avoiding possible architectural bias caused by the use of the teacher model. As opposed to the block-wise solutions discussed above, which utilize distillation as intermediate supervision, we propose a self-supervised representation learning scheme named ensemble bootstrapping to optimize each block of our supernet. To be more specific, each sampled sub-networks are trained to predict the probability ensemble of all the sampled ones in the target network, between different augmented views of the same image. In the searching stage, an unsupervised evaluation metric, is proposed to ensure fairness by searching towards the architecture population center. More specifically, the probability ensemble of all the architectures in the population is used as the evaluation target to measure the performance of the sampled models.

Additionally, we design a fabric-like hybrid CNN-transformer search space (HyTra) with searchable downsampling positions and use it as a case study for hybrid architectures to evaluate our method. In each layer of HyTra search space, CNN building blocks and transformer building blocks of different resolutions are in parallel and can be chosen flexibly. This diversified search space covers pure transformers with fixed content length and normal CNNs with progressively reduced spatial scales.

We prove that our NAS method can generalize well on three different search spaces and three datasets. On HyTra search space, our searched models outperforms the ones searched by our supervised NAS counterpart [36], proving that our method successfully avoids possible architecture bias brought by supervised distillation. Our method achieves superior architecture rating accuracy with 0.78 and 0.76 Spearman correlation on the canonical MBConv search space with ImageNet and on NATS-Bench size search space $S_2$ [16] with CIFAR-100, respectively, surpassing state-of-the-art NAS methods, proving that our method successfully suppressed the problem of inaccurate architecture rating caused by large weight-sharing space.

Our searched models on HyTra search space achieves 82.2% accuracy on ImageNet, surpassing EfficientNet [63] by 2.1%, with comparable compute time. By providing strong results through BossNet-T, we hope that this diversified HyTra search space with disparate candidates and high-performance architectures can serve as a new arena for future NAS works. We also hope that our BossNAS can serve as a widely used tool for hybrid architecture design.

2. Related Works

Block-wise weight-sharing NAS [36, 46, 83, 84] approaches factorize the supernet into independently optimized blocks and thus reduce the weight-sharing space, resolving the issue of inaccurate architecture ratings caused by weight-sharing. DNA [36] first introduced the block-wisely supervised architecture rating scheme with knowledge distillation. Based on this scheme, DONNA [46] further propose to predict an architecture rating using a linear combination of its blockwise ratings rather than a simplistic sum. SP [83] were the first to apply this scheme to network pruning. However, all of the aforementioned methods rely on a supervised distillation scheme, which inevitably introduces architectural bias from the teacher. We accordingly propose a block-wisely self-supervised scheme, which completely casts off the yoke of the teacher architecture.

Unsupervised NAS [41, 86] methods perform architecture search without access to any human-annotated labels. UnNAS [41] introduced unsupervised pretext tasks [34, 48, 85] to weight-sharing NAS for supernet training and architecture

---

2In this work, architecture rating accuracy refers to the correlation of the predicted architecture ranking and the ground truth architecture ranking.

3Following [61], compute time refers to the time spent for forward and backward passes.
Auto-DeepLab [40] presents a hierarchical search space for works directly predict the representation of one view from we explore Siamese supernets. Dilemma of NAS: efficiency or accuracy. sample-based NAS methods produce accurate architecture that can be attributed to the biased supervision, i.e. ratings, they are also computationally prohibitive. Weight-sharing rating scheme in one-shot NAS methods has brought about a tremendous reduction of search cost by encoding the entire search space $A$ into a weight-sharing supernet, with the weights $W$ shared by all the candidate architectures and optimized concurrently as: $W^* = \arg\min_{W} \mathcal{L}_{\text{train}}(W, A; x, y)$. Here $\mathcal{L}_{\text{train}}(\cdot)$ denotes the training loss function, while $x$ and $y$ denote the input data and the labels, respectively. Subsequently, architectures $\alpha$ are searched based on the ranking of their ratings with these shared network weights. Without loss of generality, we choose the evaluation loss function $\mathcal{L}_{\text{val}}$ as the rating metric; the searching phase can be formulated as: $\alpha^* = \arg\min_{\alpha} \mathcal{L}_{\text{val}}(W^*, \alpha; x, y)$. However, the architecture ranking based on the shared weights $W^*$ does not necessarily represents the correct ranking of the architectures, as the weights inherited from the supernet are highly entangled and are not fully and fairly optimized. As pointed out in the literature [57, 72, 79], weight-sharing methods suffer from low architecture rating accuracy.

**Self-supervised contrastive learning** methods [49, 70, 27, 64, 90, 23, 9] have significantly advanced the unsupervised learning of visual representations. These approaches learn visual representations in a discriminative fashion by gathering the representations of different views from the same image and spreading those from different images. Recently, the innovative BYOL [20] and SimSiam [10] learned visual representations without the use of negative examples. These works directly predict the representation of one view from another using a pair of Siamese networks with the same architectures and shared weights [10], or with one of the Siamese network branches being a momentum encoder, thereby forming a bootstrapping scheme [20]. Our work introduces a novel bootstrapping scheme with probability ensemble to our unsupervised NAS scheme to avoid the supervision bias in block-wise NAS.

**Section Search Spaces.** Cell-based search spaces, first proposed in [92], are generally used in previous NAS methods [42, 53, 43, 51] and benchmarks [73, 17, 16]. They search for a repeatable cell-level architecture, while keeping a manually designed network-level architecture. By contrast, network-level search spaces with layer-level granularity [7, 69, 14, 36, 46, 86] and block-level granularity [62, 28, 63] search for the macro network-level structure using manually designed building blocks (e.g. MBConv [55]). Auto-DeepLab [40] presents a hierarchical search space for semantic segmentation, with repeatable cells and a fabric-like network-level structure. Our HyTra search space also has a fabric-like network-level structure, albeit with layer-level granularity rather than repeated cells.

3. **Block-wisely Self-supervised NAS**

In this section, we first briefly introduce the dilemma of NAS and its block-wise solutions [36, 46, 83, 84], then present our proposed BossNAS in detail, along with its two key elements: i) unsupervised supernet training phase with ensemble bootstrapping; ii) unsupervised architecture rating and searching phase towards architecture population center.

**Notations.** We denote scalars, tensors and sets of tensors using lower case, bold lower case and upper case calligraphic letters respectively (e.g. $n$, $x$ and $\mathcal{X}$). For simplicity, we use $\{x_n\}$ to denote the set $\{x_n\}_{n=1}^m$ with cardinality $|n|$.

3.1. **Dilemma of NAS and the Block-wise Solutions**

Dilemma of NAS: efficiency or accuracy. While classical sample-based NAS methods produce accurate architecture ratings, they are also computationally prohibitive. Weight-sharing rating scheme in one-shot NAS methods has brought about a tremendous reduction of search cost by encoding the entire search space $A$ into a weight-sharing supernet, with the weights $W$ shared by all the candidate architectures and optimized concurrently as: $W^* = \arg\min_{W} \mathcal{L}_{\text{train}}(W, A; x, y)$. Here $\mathcal{L}_{\text{train}}(\cdot)$ denotes the training loss function, while $x$ and $y$ denote the input data and the labels, respectively. Subsequently, architectures $\alpha$ are searched based on the ranking of their ratings with these shared network weights. Without loss of generality, we choose the evaluation loss function $\mathcal{L}_{\text{val}}$ as the rating metric; the searching phase can be formulated as: $\alpha^* = \arg\min_{\alpha} \mathcal{L}_{\text{val}}(W^*, \alpha; x, y)$. However, the architecture ranking based on the shared weights $W^*$ does not necessarily represents the correct ranking of the architectures, as the weights inherited from the supernet are highly entangled and are not fully and fairly optimized. As pointed out in the literature [57, 72, 79], weight-sharing methods suffer from low architecture rating accuracy.

**Block-wise supervised NAS.** As proven theoretically and experimentally by [38, 36, 46], reducing the weight-sharing space (i.e. total number of weight-sharing architectures) can effectively improve the accuracy of architecture rating. In practice, block-wise solutions [36, 46, 83, 84] find a way out of this dilemma of NAS by block-wisely factorizing the search space in the depth dimension, thus reducing the weight-sharing space while maintaining the original size of the search space. Given a supernet consisting of $|k|$ blocks $S(W, A) = \{S_k(W_k, A_k)\}$, with $W = \{W_k\}$ and $A = \{A_k\}$ denoting its weights and architecture that are block-wisely separable in the depth dimension, each block of the supernet is trained separately before searching among all blocks in combination by the sum [36], or a linear combination [46], of each block’s evaluation loss $\mathcal{L}_{\text{val}}$:

$$\alpha^* = \{\alpha_k\}^* = \arg\min_{\forall\{\alpha_k\} \subset A} \sum_{k=1}^{\lfloor k \rfloor} \lambda_k \mathcal{L}_{\text{val}}(W_k^*, \alpha_k; x_k, y_k)$$

s.t. $W_k^* = \arg\min_{W_k} \mathcal{L}_{\text{train}}(W_k, A_k; x_k, y_k)$.

To isolate the training of each supernet block, given an input $x$, the intermediate input and target $\{x_k, y_k\}$ of the $k$-th block is generated by a fixed teacher network $T$ (with architecture $\alpha^T$ and ground-truth weights $W^T$): $\{x_1, y_1\} = \{x, T(x)\}$, and $\{x_k, y_k\} = \{T_k(x), T_k(x)\}$, $k > 1$, where $T_k$ represents the teacher network truncated after the $k$-th block. As the data used for both training and searching phase are generated by the teacher model $T(W^T, \alpha^T)$, the architecture ratings are likely to be highly correlated with the teacher architecture. For instance, a convolutional teacher have a limited receptive field and distinctive architectural inductive biases like translation equivariance. With such a biased supervision, candidate architectures are likely to be trained and rated unfairly. We observe two phenomena that can be attributed to the biased supervision, i.e. candidate
Candidates

Online

Supernet

Blocks

Downsample & MLPs

Feature Flow

Best Path

E Ensemble Module

l2

Distance

Stop

Gradient

Online Supernet

EMA

Supernet

l2

EMA Supernet

E

Block k

Supernet

Figure 2: Illustration of the Siamese supernets training with ensemble bootstrapping.

**preference and teacher preference.** Detailed experimental analysis of these two phenomena is provided in Sec. 5. To break these restrictions of current block-wise NAS solutions, we explore a scheme without using a teacher model.

### 3.2. Training with Ensemble Bootstrapping

Starting from the dual network scheme with student-teacher pair \( \{ S(W, A), T(W^T, \alpha^T) \} \), the first step to cast off the yoke of the teacher architecture is to assign \( \alpha^T = A \), thus forming a pair of Siamese supernets.

**Bootstrapping with Siamese Supernets.** To optimize such Siamese networks block-wisely, we adopt a self-supervised contrastive learning scheme. More specifically, these two supernets receive a pair of augmented views \( \{ x_1, x_2 \} \) of the same training sample \( x \) and generate the outputs \( \{ S(W, A; x_1), T(W^T, A; x_2) \} \), respectively. Analogous to previous teacher-student settings, the Siamese supernets are optimized by minimizing the distance between their outputs. In previous Siamese networks and self-supervised contrastive learning methods, the two networks either share their weights [9, 10] (i.e. \( W^T = W \)) or form a mean teacher scheme with Exponential Moving Average (EMA) \[23, 20\] (i.e. \( W^T = W^* \), where \( W^* = \tau W_{t-1} + (1 - \tau) W \)) represents the temporal average of \( W \), with \( t \) being a training timestamp, and \( \tau \) denoting the momentum factor that controls the updating speed of \( W^* \). By learning representation from the mean teacher, analogous to the simple yet powerful BYOL [20], our supernet can be optimized in an unsupervised manner without relying on a fully supervised teacher network:

\[
W^*_k = \arg \min_{W_k} \mathcal{L}_{\text{train}} \left( \{ W_k, W_k^* \}, A_k ; x_k \right). \tag{2}
\]

To eliminate the influence of pixel-wise differences between two intermediate representations caused by augmentations (e.g. random crop), as well as to ensure better generalization on candidate architectures with different receptive fields or even different resolutions, we project the representations to the latent space before calculating the element-wise distance.

**Ensemble Bootstrapping.** However, unlike single networks, supernets are typically optimized by path sampling strategies, e.g. single path [21] or fair path [14]. When naively adopting bootstrapping, each sub-network learns from the moving average of itself. In the absence of a common objective, the weights shared by different sub-networks suffer from convergence hardness, leading to training instability and inaccurate architecture ratings. To address this problem, we propose an unsupervised supernet training scheme, named ensemble bootstrapping.

Considering \(|p|\) sub-networks \( \{ \alpha_p \} \subset A_k \) sampled from the \( k \)-th block of the search space \( A \) in the \( t \)-th training iteration, and given a training sample \( x \), \(|p|\) pairs of augmented views \( \{ x_p \} \sim p_{\text{aug}}(\cdot | x) \), \( \{ x'_p \} \sim p_{\text{aug}}(\cdot | x) \) are generated for each sampled sub-network of the Siamese supernets. To form a common objective for all paths, we can use a scheme analogous to ensemble distillation \[58, 59\] in supervised learning. As illustrated in Fig. 2, each sampled sub-network of the online supernet learns to predict the probability ensemble of all sampled sub-networks in the EMA supernet:

\[
\mathbb{T}_k (\{ \alpha_p \}; \{ x'_p \}) = \frac{1}{|p|} \sum_{p=1}^{|p|} \mathbb{T}_k (W^*_p, \alpha_p ; x'_p). \tag{3}
\]

In summary, the block-wisely self-supervised training process of the Siamese supernets is formulated as follows:

\[
W^*_k = \arg \min_{W_k} \sum_{p=1}^{|p|} \mathcal{L}_{\text{train}} \left( \{ W_k, W_k^* \}, \{ \alpha_p \}; x \right),
\]

where

\[
\mathcal{L}_{\text{train}} \left( \{ W_k, W_k^* \}, \{ \alpha_p \}; x \right) = \left| \sum_{p=1}^{|p|} \mathbb{T}_k (W^*_p, \{ \alpha_p \}; x) \right|^2.
\]

### 3.3. Searching Towards the Population Center

After the convergence of the Siamese supernets is complete, the architectures can be ranked and searched by the rating determined based on the weight of the supernets, as in Eqn. 1. In this section, we design a fair and effective unsupervised rating metric \( \mathcal{L}_{\text{val}} \) for searching phase.

To evaluate the performance of a network trained with contrastive self-supervision, previous works [23, 9, 20, 10] have utilized supervised metrics, such as accuracies of linear evaluation or few-shot classification. To develop an unsupervised NAS method, we aim to avoid schemes that depend on human-annotated labels and instead pursue a completely unsupervised evaluation metric. Previous unsupervised NAS methods [41, 86] utilize either the accuracy of pretext tasks or convergence measurement with angle-based metrics to rate candidate architectures. Unfortunately, the losses of self-supervised contrastive learning do not necessarily represent either the architecture performance or the architecture
convergence, as the input views and target networks are both randomly sampled. Moreover, the target networks are somewhat biased and cannot serve as ground truth targets. To avoid these concerns, we propose a fair and effective unsupervised evaluation metric for architecture search.

Without loss of generality, we consider searching with an evolutionary algorithm [11, 53], where architectures are optimized by evolving an architecture population \( \{ \alpha_p \} \). Analogous to the optimization of the weights, we propose to use probability ensemble among the population \( \{ \alpha_p \} \) as the common target to provide a fair rating for each architecture \( \alpha_p \). Additionally, one pair of views \( \{ x_1, x_2 \} \) for each validation sample \( x \) are generated and fixed to avoid the bias introduced by variable augmentation. In parallel to Eqn. 3, we have the probability ensemble of the architecture population:

\[
\hat{S}_k(\{ \alpha_p \}; x_2) = \frac{1}{|P|} \sum_{p=1}^{|P|} \hat{S}_k(\alpha_p; x_2).
\] (5)

In practice, by dividing the super net into medium-sized blocks (e.g. 4 layers of 4 candidates, \( 4^4 = 256 \) architectures), traversal evaluation of all the candidate architectures are affordable. In this case, the architecture population \( \{ \alpha_p \} \) is expanded to the whole block-wise search space \( A_k \), and the whole searching process is finished in a single step:

\[
\alpha^* = \arg \min_{\alpha \in \mathcal{A}} \sum_{k=1}^{k_{\text{max}}} \lambda_k \mathcal{L}_{ul}(\alpha; x_k)
\] (6)

where

\[
\mathcal{L}_{ul}(\alpha; x) = \| \hat{S}_k(\alpha; x_1) - \hat{S}_k(A_k; x_2) \|^2_2.
\]

4. Hybrid CNN-transformer Search Space

In this section, we present a fabric-like hybrid CNN-transformer search space, named HyTra, with disparate candidate building blocks and flexible down-sampling positions.

4.1. CNN and Transformer Candidate Blocks

The first step in designing a hybrid CNN-transformer search space is to include the proper CNN and transformer building blocks. These two types of building blocks should be able to perform well either when simply aggregated in sequence or when combined freely. We choose the classical and robust residual bottleneck (ResConv) in ResNet [24] as the CNN candidate building block. In parallel, we design a lightweight and robust transformer building block ResAtt based on the pluggable BoTBlock [61] and NLBlock [67].

Computation Balancing with Implicit Position Encodings. To facilitate fair and meaningful competition, candidate building blocks should have similar computation complexities. The original BoTBlock is slower than ResConv, as its relative position encodings are computed separately through multiplication with the query. Simply removing the content-position branch from BoTBlocks, resembling to NLBlocks, could reduce their compute time to make them comparable to ResConv. However, position encodings are crucial for vision transformers to achieve good performance. In CPVT [13], the authors uses single convolutions in between transformer encoder blocks as the position encoding generator. Similarly, we replace the relative position encoding branch in BoTBlock with a light depthwise separable convolution as an implicit position encoding module, forming our ResAtt. By this simple modification, we reduce the computation complexity of position encoding module from \( \mathcal{O}(CW^3) \) to \( \mathcal{O}(W^2) \), with \( C \) denoting number of channels and \( W \) denoting the width or height. In contrast to CPVT and BoT (Fig. 4), our position encoding modules (Fig. 3 right) are placed between the input projection layer and the self-attention module. In addition, our implicit position encoding modules are also responsible for down-sampling. This modification is also applied to ResConv, which enables weight sharing between candidate blocks with different down-sampling rates (i.e. 1 or 2).

4.2. Fabric of Hybrid CNN-transformers

Beyond the building blocks, CNNs and transformers differ considerably in terms of their macro architectures. Unlike CNNs, which process images in stages with various spatial sizes, transformers typically do not change sequence length (image patches) and retains the same scale at each layer. As shown in Fig. 3 left, to cover both the CNNs and transformers, our search space is designed with flexible down-sampling positions, forming a fabric [56] of Hybrid CNN-transformers. At each choice block layer of the fabric, the spatial resolution can either stay unchanged or be reduced to half of its scale, until reaching the smallest scale. This fabric-like search space contains architectures resembling the popular vision transformers [19, 65, 13], CNNs [24, 30] and hybrid CNN-transformers [61] at different scales.
Table 1: ImageNet results of state-of-the-art models and our searched hybrid CNN-transformers. Compute steptime is measured on a single GeForce RTX 3090 GPU with batch size 32. Purple is used to denote manually selected architectures from search space HyTra. †: Directly tested on larger input size without finetuning (i.e. 288 for BossNet-T0† and 256 for BossNet-T1†).

5. Experiments

Setups. We evaluate our method on three search spaces, including our proposed HyTra search space and other two existing search spaces, i.e. MBConv search space [7, 36] and NATS-Bench size search space $S_5$ [16]. The datasets we use to evaluate and analyze our method are ImageNet [15], CIFAR-10 and CIFAR-100 [35]. We train each block size \( S \) [16]. The datasets measured on a single GeForce RTX 3090 GPU with batch size 32.

5.1. Searching for Hybrid CNN-transfomer

Analyze of HyTra search space. We manually stitched four architectures on our fabric-like HyTra search space, following as closely as possible to previous human-designed networks [24, 19, 61, except using our \{ResConv, ResAtt\} building blocks. As shown in Tab. 1, these models (in purple) consistently outperform their prototypes. Remarkably, BoT50-T surpasses the original BoT50 by 1.2% top-1 accuracy with 1.17× compute time reduction, proving the superiority of our designed building blocks.

Performance of searched models. With our proposed search space and NAS method, we explore hybrid CNN-transformer architectures on ImageNet. The results of our searched models (BossNet-T) and models with comparable compute time are summarized in Tab. 1.

Firstly, BossNet-T0 outperforms a wide range of state-of-the-art models. For instance, BossNet-T0 without SE module achieves 80.5% top-1 accuracy, surpassing the human-designed hybrid CNN-transformer, BoTNet50, by 2.2% while being 1.19× faster in terms of compute time; when equipped with SE and SiLU activation, BossNet-T0 further achieves 80.8% top-1 accuracy, surpassing the NAS searched EfficientNet-B1 by 1.7× while being 1.14× faster.

Secondly, our searched model demonstrates absolute superiority over manually and randomly selected models from search space HyTra. In particular, BossNet-T0 achieves up to 6.0% improvement over manually selected models, proving the effectiveness of our architecture search.

Thirdly, BossNet-T0 outperforms other recent NAS methods on search space HyTra. BossNet-T0 achieves 0.5% accuracy gains over DNA-T, which is searched by our supervised NAS counterpart [36].

Finally, when extended to larger model size and input size, the family of BossNet-T models maintain its superiority. By removing the downsampling in the last stage of BossNet-T0 (same scheme as BoTNet-S1 [61]), we have BossNet-T1, which achieves 81.9% accuracy, surpassing EfficientNet-B2 by 1.8%. By directly testing on larger input resolutions without finetuning, BossNet-T0† (on 288 × 288 input size) achieves 81.6% top-1 accuracy, and outperforms BoTNet50 + SE by 2.0% ; BossNet-T1† (on 256 × 256 input size) achieves 82.2% top-1 accuracy, surpassing T2T-ViT-19 and EfficientNet-B2 by 1.0% and 2.1%, respectively. Note that typical transformers or hybrid CNN-transformers with explicit position encodings cannot be directly transferred to larger input size without finetuning.

Architecture visualization and analysis. We visualize the architecture of DNA-T and BossNet-T0 in Fig. 5. DNA-T clearly prefers convolutions, as it contains 13 ResConv blocks and only three ResAtt blocks. By contrast, BossNet-T0 has similar numbers of convolutions and attentions and eventually achieves a higher accuracy. We refer this to Phenomenon I: candidate preference, and attribute it to archi-
As BossNAS performs traversal search (i.e., accuracy of searching phase is 100%), the architecture rating accuracy directly represents its effectiveness. We use the 23 open-sourced architectures in MBConv search space and their corresponding ground truth accuracies provided by [36] to calculate the architecture rating accuracy, i.e., the ranking correlation between the predicted architecture ranking and the ground truth model ranking. We use three different ranking correlation metrics:

- Kendall Tau (τ)
- Spearman Rho (ρ)
- Pearson R (R)

### Table 2: ImageNet results of state-of-the-art NAS models on MBConv search space.

| Method          | Search Cost | τ       | ρ       | R       |
|-----------------|-------------|---------|---------|---------|
| SPOS [21]       | 8.5 Gds     | -0.18   | -0.27   | -0.29   |
| DARTS [43]      | 50 Gds      | 0.08    | 0.14    | 0.06    |
| MnasNet [62]    | 288 Tds     | 0.61    | 0.77    | 0.78    |
| DNA [36] (EffNetB0) | 8.5 Gds | 0.62    | 0.77    | 0.83    |
| DNA [36] (MBNetV1) | 8.5 Gds | 0.23    | 0.27    | 0.37    |
| BossNAS         | 10 Gds      | 0.65    | 0.78    | 0.85    |

### Table 3: Comparison of the effectiveness and efficiency of different NAS methods on MBConv search space and ImageNet dataset.

| Method              | MAdds (M) | Top-1 (%) | Top-5 (%) | Unsupv. | Supv. |
|---------------------|-----------|-----------|-----------|---------|-------|
| FairNAS [14]        | 388M      | 75.3      | 92.4      | 76.2    | 77.1  |
| ProxylessNAS [7]    | 465M      | 75.1      | 92.5      | 70.7    | 71.0  |
| FBNet-C [69]        | 375M      | 74.9      | -         | 69.5    | 70.7  |
| SPOS [21]           | 472M      | 74.8      | -         | 76.2    | 77.1  |
| RLNAS [86]          | 473M      | 75.6      | 92.6      | 76.2    | 77.1  |
| BossNet-M1 w/o SE  | 475M      | 76.2      | 93.0      | 76.2    | 77.1  |

| Method              | C-10 (%)  | C-100 (%) | τ       | ρ       | R       |
|---------------------|-----------|-----------|---------|---------|---------|
| FBNet v2 [66]       | 93.15     | 70.72     | -       | -       | -       |
| TuNAS [5]           | 92.78     | 70.11     | -       | -       | -       |
| CE [26]             | 90.55     | 70.78     | 0.43    | 0.60    | 0.60    |
| BossNAS             | 93.29     | 70.86     | 0.59    | 0.76    | 0.79    |

### Table 4: Comparison of searched model accuracy and architecture rating accuracy of different NAS methods on NATS-Bench S₆ (C-10: CIFAR-10, C-100: CIFAR-100).

| Training         | Evaluation | τ       | ρ       | R       |
|------------------|------------|---------|---------|---------|
| Supv. distill.   | Supv. distill. | 0.62    | 0.77    | 0.83    |
| Supv. class.     | Supv. class. | 0.46    | 0.65    | 0.71    |
| Unsupv. bootstrap| Unsupv. eval | 0.12    | 0.15    | 0.28    |
| Unsupv. EB       | Supv. linear eval | 0.55    | 0.73    | 0.79    |
| Unsupv. EB       | Unsupv. eval | 0.65    | 0.78    | 0.85    |

### Table 5: Ablation analysis of training methods and evaluation methods on MBConv Search Space.

Kendall Tau (τ) [33], Spearman Rho (ρ) and Pearson R (R). All three metrics range from -1 to 1, with “-1” representing a completely reversed ranking, “1” meaning an entirely correct ranking, and “0” representing no correlation between rankings. As shown in Tab. 3 and Fig. 6 left, our BossNAS achieves high rating accuracies with massive search cost. BossNAS successfully addressed such dilemma of NAS by achieving even higher rating accuracies than MnasNet (e.g. 0.07 R) with 28.8× acceleration in search cost. Second, supervised block-wise NAS method, DNA [36] fails to achieve high rating accuracies when using a teacher that are largely different from the candidate architecture (MobileNetV1 [29] vs. EfficientNet-based candidates [63]), which we refer to as Phenomenon II: teacher preference. Our unsupervised BossNAS achieves higher rating accuracies than DNA (0.03 τ), successfully casting off the yoke of the teacher network.

### 5.3. Results on NATS-Bench S₆

For NATS-Bench size search space S₆, experiments are conducted on two datasets: CIFAR-10 and CIFAR-100.
effectiveness of our BossNAS.

In addition, the architecture rating accuracies on CIFAR-100 dataset are shown in Tab. 4. Our method, without access to the ground truth architecture accuracies and even without access to any human-annotated labels, outperforms a predictor-based NAS method [26], which is trained with ground truth architecture accuracies, by a large gap (i.e. 0.16 $\tau$ and 0.19 $R$). More analysis on NATS-Bench $S_S$ could be found in Appendix A.3.

5.4. Ablation Study

We perform extensive ablation studies to analyze our proposed training and evaluation methods separately in this section. All experiments are conducted on MBConv search space with ImageNet. Note that the ablation analysis of the HyTra search space has already been presented in Sec. 5.1.

**training methods.** We compared several training methods for the block-wise supernet: (1) *Supervised distillation* method ($Supv$. distill.), using a pre-trained teacher model to provide block-wise supervision, i.e. the training scheme used in DNA [36] (2) *Supervised classification* ($Supv$. class.), using real labels directly as the block-wise supervision. (3) *Unsupervised bootstrapping* ($Unsupv$. bootstrap), where the Siamese supernets are optimized by bootstrapping the corresponding paths in the two networks. (4) Our *unsupervised ensemble bootstrapping* method ($Unsupv$. EB), where each sampled paths are optimized by learning to predict the probability ensemble of sampled paths from the mean teacher. As shown in Tab. 5, our training method surpasses all others, achieving the best results in architecture rating accuracy. In particular, by comparing $Unsupv$. bootstrap with $Unsupv$. EB in the third and fifth line, we can see that replacing our proposed EB with the naive bootstrapping scheme, the architecture rating accuracy drops sharply by 0.53 $\tau$. Without the probability ensemble, bootstrapping fails to reach a reasonable rating accuracy, proving that the proposed ensemble bootstrapping is indispensable for our BossNAS.

**Evaluation methods.** Similar to the ablation analysis of training methods, we compare our evaluation methods with (1) *Supervised distillation* method ($Supv$. distill.) and (2) *Supervised classification* ($Supv$. class.). Additionally, to perform ablation analysis of evaluation without changing our unsupervised training method, we also compare with (3) supervised linear evaluation ($Supv$. linear eval), where architectures are rated by fixing the weights of the supernet and finetuning a weight sharing linear classifier to evaluate each architecture. (4) Our unsupervised evaluation metric

![Figure 6: Left: Ranking correlations of 6 different NAS methods on MBConv Search Space. Right: Architecture ranking of BossNAS on NATS-Bench $S_S$. In all the diagrams, x-axis denotes ground truth accuracy; y-axis denotes evaluation metrics.](Image)

![Figure 7: Ranking correlations during supernet training. (Unsupv. eval) that rate architectures by its distance to the ensemble probability center of the whole searching space. From the last two rows of Tab. 5, we surprisingly found that our $Unsupv$. eval outperforms supervised linear evaluation scheme in architecture rating by a remarkable gap (0.1 $\tau$).](Image)

5.5. Convergence Behavior

To further demonstrate the effectiveness of BossNAS, we investigate the architecture rating accuracy during the supernet training process on MBConv search space with ImageNet. The three ranking correlation metrics of our BossNAS during its 20 training epochs are shown in Fig. 7. The architecture rating accuracy increases rapidly in the early stage and continues to grow with minor fluctuation. The rating accuracy converges at the 12-th epoch and continues to be stable till the end of the training phase. The stably increasing architecture rating ability proves the stability of our BossNAS. In addition, the fast converging ranking correlation demonstrates that our method is easy to optimize and do not require longer training. Please refer to Appendix A.3 for analysis of convergence behavior on NATS-Bench $S_S$.

6. Conclusion

In this work, we present BossNAS, a general, unsupervised NAS method with a self-supervised training technique named ensemble bootstrapping and an unsupervised evaluation metric for architectures. Experiments on three search spaces prove that our method successfully addressed the problem of inaccurate architecture rating caused by large weight-sharing space while avoiding the architectural bias brought by supervised distillation. Ablation analysis proved that the two components, ensemble bootstrapping scheme and unsupervised evaluation metric, are both crucial for our method. Additionally, we present a fabric-like search space named HyTra. On this challenging search space, our searched hybrid CNN-transformer model, achieves 82.2% accuracy on ImageNet, surpassing EfficientNet by 2.1% with comparable compute time.

**Acknowledgement**

Dr Xiaojun Chang is partially supported by Australian Research Council (ARC) Discovery Early Career Research Award (DECR A) under grant no. DE190100626.
Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In CVPR, 2018.

Shen Yan, Yu Zheng, Wei Ao, Xiao Zeng, and Mi Zhang. Does unsupervised architecture representation learning help neural architecture search? arXiv:2006.06936, 2020.

Antoine Yang, Pedro M. Esperança, and Fabio Maria Carlucci. NAS evaluation is frustratingly hard. In ICLR, 2020.

Chris Ying, Aaron Klein, Eric Christiansen, Esteban Real, Kevin Murphy, and Frank Hutter. Nas-bench-101: Towards reproducible neural architecture search. In ICML, 2019.

Yang You, Igor Gitman, and Boris Ginsburg. Scaling sgd batch size to 32k for imagenet training. arXiv:1708.03888, 2017.

Jiahui Yu and Thomas Huang. Autoslim: Towards one-shot architecture search for channel numbers. arXiv:1903.11728, 2019.

Jiahui Yu and Thomas S. Huang. Universally slimmable networks and improved training techniques. In ICCV, 2019.

Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas S. Huang. Slimmable neural networks. In ICLR, 2019.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. arXiv:2101.11986, 2021.

Arber Zela, Julien Siems, and Frank Hutter. Nas-bench-1shot1: Benchmarking and dissecting one-shot neural architecture search. In ICLR, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. arXiv:1611.03530, 2016.

Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, Zongyang Ge, and Steven W. Su. Differentiable neural architecture search in equivalent space with exploration enhancement. In NeurIPS, 2020.

Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, and Steven W. Su. Overcoming multi-model forgetting in one-shot NAS with diversity maximization. In CVPR, 2020.

Mingyang Zhang and Linlin Ou. Stage-wise channel pruning for model compression. arXiv:2011.04908, 2020.

Man Zhang, Yong Zhou, Jiaqi Zhao, Shixiong Xia, Jiaqi Wang, and Zizheng Huang. Semi-supervised blockwisely architecture search for efficient lightweight generative adversarial network. Pattern Recognition, 112:107794, 2021.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In ECCV, 2016.

Xuanyang Zhang, Pengfei Hou, Xiangyu Zhang, and Jian Sun. Neural architecture search with random labels. arXiv:2101.11834, 2021.

Sixiaoz Zheng, Jiachen Lu, Hengshuang Zhao, Xiadian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. arXiv:2012.15840, 2020.
A. Appendix

A.1. A brief review of NAS

NAS methods aim to automatically optimize neural network architectures by exploring search spaces with search algorithms and evaluating architectures by means of rating schemes. NAS methods can be divided into two categories depending on the rating scheme utilized, i.e. multi-trial NAS and weight-sharing NAS. Multi-trial NAS methods [91, 2, 53, 62, 42, 82] rate all sampled architectures by training them from scratch, making this process computationally prohibitive and difficult to deploy on large datasets. They either perform architecture rating by training on relatively small datasets (e.g. CIFAR-10) [91, 2, 53] or by training for the first few epochs (e.g. 5 epochs) [62] on ImageNet. To avoid repeated training of candidate networks, weight-sharing NAS methods [7, 43, 18, 1, 6, 12, 81] optimize the weight of the supernet and the factors (or agent) from the supernet. Among them, gradient-based approaches [43, 7, 69] and sampler-based approaches [51, 60] jointly optimize the weight of the supernet and the factors (or agent) used to choose the architecture; for their part, one-shot approaches [21, 14, 6, 4] optimize the supernet before performing a search with the frozen supernet weights. We refer to [54] for a more comprehensive NAS review.

A.2. Implementation Details

Search spaces. We evaluate our method on three search spaces:

- **HyTra search space.** The beginning of the networks in this search space is the classic ResNet stem that reduces the spatial resolution by a factor of 4 with a strided 7×7 convolution layer and a max-pooling layer. It contains \( L = 16 \) choice block layers in total, as the same to ResNet50. Before the first choice block layer, the input can be further down-sampled to different scales. The downsampling module consists of multiple 3×3 convolutions with stride of 2. At each choice block layer, the spatial resolution can either stay unchanged or be reduced to half of its scale, unless reaching the smallest scale \( 1/32 \). As introduced in Sec. 4, this search space contains two disparate candidate choices: \{ResConv, ResAtt\}. As transformer blocks are expensive in the first scales, we only enable the choice of ResAtt in the last two scales (i.e. 1/16 and 1/32). The total size of this challenging hybrid search space is roughly \( 2.8 \times 10^6 \).

- **MBConv search space.** MobileNet-like search space and its variations are generally used as benchmarks for recent NAS methods [62, 28, 63, 7, 69, 14, 36, 46, 86]. Following Li et al. [36], we use a search space with 18 layers and each layer contains 4 candidate MobileNet blocks (combination of kernel size \( \{3, 5\} \) and reduction rate \( \{3, 6\} \)). This results in a large search space containing about \( 4^{18} \approx 6.9 \times 10^{10} \) architectures.

- **NATS-Bench \( S_S \).** The NATS-Bench size search space \( S_S \) [16] is a channel configuration search space built upon a fixed cell-based architecture with 5 layers, where the 2-nd and 4-th layers have a down-sample rate of 2. Number of channels in each layer is chosen from \{8, 16, 24, 32, 40, 48, 56, 64\}. \( S_S \) has \( 8^5 = 32768 \) architecture candidates in total. Candidates of different channel numbers in our supernet share the weights in a slimmable manner [77, 76, 75, 37]. We divide the supernet into 3 blocks, according to spatial size.

Datasets. The datasets we use to evaluate and analysis our method include ImageNet [15], CIFAR-10 and CIFAR-100 [35]. ImageNet is a large-scale dataset containing 1.2 M train set images and 50 K val set images in 1000 classes. We randomly samples 50 K images from the original train set to form a NAS-val set for architecture rating and use the remainder as the NAS-train set for supernet training. No labels are used during training and searching of our NAS method. Finally, our searched architectures are retrained from scratch on train set and evaluated on val set. For CIFAR-10 and CIFAR-100 [35], we use the splits proposed in NATS-Bench [16]. CIFAR-10 is divided into 25 K train set, 25 K val set, and 10 K test set. CIFAR-100 is divided into 50 K train set, 5 K val set, and 5 K test set. The final accuracies of searched architectures are queried from NATS-Bench \( S_S \) [16].

Training details.

We train each block of the BossNAS supernet for 20 epochs including 1 linear warm-up epoch on ImageNet. For the relatively smaller CIFAR datasets, we extend it to 30 epochs. In each training step, we randomly sample 4 paths for the ensemble bootstrapping. Other hyperparameters for self-supervised training of the supernet follow closely to BYOL [20], we use the LARS optimizer [74] with a cosine decay learning rate schedule [44]. The base learning rate is set to 4.8 for a total batchsize of 4096.

For ImageNet retraining of BossNet-T models, we follow strictly the same with DeiT [65] without further tuning, as we found it robust for both CNNs and transformers. More specifically, we use AdamW optimizer with \( 1 \times 10^{-3} \) initial learning rate and cosine learning rate scheduler, for a total batch size of 1024. Weight decay is set to 0.05. Please refer to DeiT [65] for more details on data-augmentation and regularization.
| Dataset    | Method      | $\tau$ | $\rho$ | $R$  |
|------------|-------------|--------|--------|------|
| CIFAR-10   | CE [26]     | 0.42   | 0.60   | 0.59 |
|            | BossNAS     | 0.53   | 0.73   | 0.72 |
| CIFAR-100  | CE [26]     | 0.43   | 0.60   | 0.60 |
|            | BossNAS     | 0.59   | 0.76   | 0.79 |

Table 6: Architecture rating accuracy on NATS-Bench $S_S$ with CIFAR datasets.

Table 6: Architecture rating accuracy on NATS-Bench $S_S$ with CIFAR datasets.

For ImageNet retraining of BossNet-M models, we follow closely to EfficientNet [63]. We use batchsize 4096, RMSprop optimizer with momentum 0.9 and initial learning rate of 0.256 which decays by 0.97 every 2.4 epochs. Please refer to EfficientNet [63] for more details of other settings.

Re-implementation of other NAS methods on HyTra.

For DNA [36], we use ResNet-50 [24] as the teacher model. We divide the supernet into four blocks, with four layers in each block, and train each block for 20 epochs. The intermediate features of every block of the student supernet and the teacher are all downsampled with global pooling and projected with one fully-connected layer before calculating distillation loss, as the scale of different candidate block is not the same in HyTra search space. Other settings follow closely to DNA [36].

For UnNAS [41], we adopt rotation prediction [34] ($\text{Rot}$) pretext task, for its simplicity. Following [41], we use three extra stride-2 convolution layers at the beginning of the supernet to reduce spatial resolution. The supernet is trained for 2 epochs as in [41].

A.3. Additional Analysis on NATS-Bench $S_S$

Architecture rating comparison. We compare with the predictor-based NAS method CE [26] by architecture rating accuracy on CIFAR-10 and CIFAR-100. As shown in Fig. 8, we compare the two NAS methods by plotting the correlation of the architecture rating and the true accuracy of 3000 randomly sampled architectures from NATS-Bench size search space $S_S$ [16]. Architectures with BossNAS form denser and more spindly scatter pattern than CE on both of the two datasets. Moreover, as measured quantitatively in Tab. 6, BossNAS outperforms CE by a large margin (0.11 and 0.16 $\tau$) in both datasets.

Convergence Behavior. We illustrate the architecture rating accuracy of BossNAS during its 30 epoch supernet training phase on CIFAR datasets in Fig. 9. The architecture rating accuracy increases quickly and steadily with minor fluctuations, in a similar manner with that on MBConv search space (Fig. 7). In particular, architecture rating accuracy of our BossNAS converges to a satisfactory result, 0.76 $\rho$, smoothly and quickly within only 20 epochs on CIFAR-100, and continues to be stable for the subsequent 10 epochs.

Figure 8: Comparison of architecture rating and its true accuracy of our BossNAS and CE [26] on NATS-Bench $S_S$ with CIFAR datasets.

Figure 9: Convergence behavior of BossNAS on NATS-Bench $S_S$ and CIFAR datasets.