

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

Changlin Li¹, Tao Tang², Guangrun Wang^{3,4}, Jiefeng Peng³, Bing Wang⁵, Xiaodan Liang^{2*}, Xiaojun Chang⁶

¹GORSE Lab, Dept. of DSAI, Monash University ²Sun Yat-sen University ³DarkMatter AI Research

⁴University of Oxford ⁵Alibaba Group ⁶RMIT 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 **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 down-sampling positions. On this challenging search space, our searched model, *BossNet-T*, achieves up to 82.5% accuracy on ImageNet, surpassing EfficientNet by 2.4% with comparable compute time. Moreover, our method achieves superior architecture rating accuracy with 0.78 and 0.76 Spearman correlation on the canonical MBCConv search space with ImageNet and on NATS-Bench size search space with CIFAR-100, respectively, surpassing state-of-the-art NAS methods.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

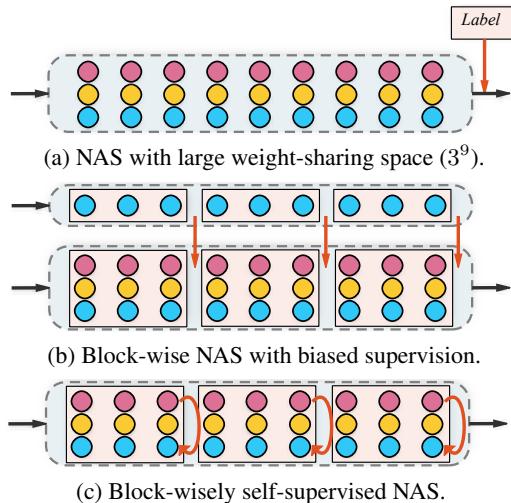


Figure 1: Comparision of three NAS schemes. Red arrows represent the supervision during training and searching.

examples of such models include ResNet [25], SENet [31], MobileNet [30] and EfficientNet [64]. 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 [20] and DETR [8], are achieving competitive performance on various vision tasks, such as image classification [20, 66, 79, 14, 9], object detection [8, 90, 62], semantic segmentation [88], and others [26, 50, 33]. As suggested by prior works [20, 62, 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

*Corresponding Author.

¹Code: <https://github.com/changlin31/BossNAS>.

architectures in a predefined search space. Representative success in performing NAS on manually designed building blocks include MobileNetV3 [29], EfficientNet [64], etc. These works are searched by *multi-trial* NAS methods [63, 92, 2, 89, 12, 47], which are computationally prohibitive (costing thousands of GPU days). Recent *weight-sharing* NAS methods [6, 53, 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 [37, 39]) as the main culprit of inaccurate architecture rating² in weight-sharing NAS methods. To reduce the size of the large weight-sharing space, previous works [37, 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 [87]) 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-wisely 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

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

performance of the sampled models.

Additionally, we design a fabric-like **hybrid CNN-transformer** search space (**HyTra**) with searchable down-sampling 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 [37], 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_S [17] 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 82.5% accuracy on ImageNet, surpassing EfficientNet [64] by 2.4%, with comparable compute time³. 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 [37, 46, 84, 85] 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 [37] 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 [84] 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, 87] methods perform architecture search without access to any human-annotated labels. UN-NAS [41] introduced unsupervised *pretext tasks* [35, 48, 86] to weight-sharing NAS for supernet training and architecture

³Following [62], *compute time* refers to the time spent for forward and backward passes.

rating. RLNAS [87] optimized the supernet using *random labels* [81, 45] and further rated architectures by means of a *convergence-based angle* metric [32]. Another line of NAS methods [72, 69, 27, 51] belonging to the category of *supervised NAS* perform unsupervised pretraining of network accuracy predictor or supernet before *supervised* finetuning or evaluation. Differing from aforementioned works in motivation and methodology, we explore *self-supervised contrastive learning* methods in our unsupervised NAS scheme to avoid the supervision bias in block-wise NAS.

Self-supervised contrastive learning methods [49, 71, 28, 65, 91, 24, 10] 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 [21] and SimSiam [11] 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 [11], or with one of the Siamese network branches being a momentum encoder, thereby forming a bootstrapping scheme [21]. Our work introduces a novel bootstrapping scheme with probability ensemble to *Siamese supernets*.

Architecture Search Spaces. Cell-based search spaces, first proposed in [93], are generally used in previous NAS methods [42, 54, 43, 52] and benchmarks [74, 19, 17]. 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, 70, 15, 37, 46, 87] and block-level granularity [63, 29, 64] search for the macro network-level structure using manually designed building blocks (*e.g.* MBCConv [56]). Auto-DeepLab [40] presents a hierarchical search space for semantic segmentation, with repeatable cells and a fabric-like [57] 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 [37, 46, 84, 85], 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 , \mathbf{x} and \mathcal{X}). For simplicity, we use $\{\mathbf{x}_n\}$ to denote the set $\{\mathbf{x}_n\}_{n=1}^{|n|}$ 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 \mathcal{A} into a weight-sharing supernet, with the weights \mathcal{W} shared by all the candidate architectures and optimized concurrently as: $\mathcal{W}^* = \arg \min_{\mathcal{W}} \mathcal{L}_{\text{train}}(\mathcal{W}, \mathcal{A}; \mathbf{x}, \mathbf{y})$. Here $\mathcal{L}_{\text{train}}(\cdot)$ denotes the training loss function, while \mathbf{x} and \mathbf{y} denote the input data and the labels, respectively. Subsequently, architectures α 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}_{val} as the rating metric; the searching phase can be formulated as: $\alpha^* = \arg \min_{\forall \alpha \in \mathcal{A}} \mathcal{L}_{\text{val}}(\mathcal{W}^*, \alpha; \mathbf{x}, \mathbf{y})$. However, the architecture ranking based on the shared weights \mathcal{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 [58, 73, 80], weight-sharing methods suffer from low architecture rating accuracy.

Block-wisely supervised NAS. As proven theoretically and experimentally by [39, 37, 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 [37, 46, 84, 85] 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 $\mathcal{S}(\mathcal{W}, \mathcal{A}) = \{\mathcal{S}_k(\mathcal{W}_k, \mathcal{A}_k)\}$, with $\mathcal{W} = \{\mathcal{W}_k\}$ and $\mathcal{A} = \{\mathcal{A}_k\}$ 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 [37], or a linear combination [46] (with weights $\{\lambda_k\}$), of each block’s evaluation loss \mathcal{L}_{val} :

$$\alpha^* = \{\alpha_k\}^* = \arg \min_{\forall \{\alpha_k\} \subset \mathcal{A}} \sum_{k=1}^{|k|} \lambda_k \mathcal{L}_{\text{val}}(\mathcal{W}_k^*, \alpha_k; \mathbf{x}_k, \mathbf{y}_k). \quad (1)$$

$$s.t. \quad \mathcal{W}_k^* = \arg \min_{\mathcal{W}_k} \mathcal{L}_{\text{train}}(\mathcal{W}_k, \mathcal{A}_k; \mathbf{x}_k, \mathbf{y}_k)$$

To isolate the training of each supernet block, given an input \mathbf{x} , the intermediate input and target $\{\mathbf{x}_k, \mathbf{y}_k\}$ of the k -th block is generated by a fixed teacher network \mathcal{T} (with architecture α^T and ground-truth weights \mathcal{W}^T): $\{\mathbf{x}_1, \mathbf{y}_1\} = \{\mathbf{x}, \mathcal{T}_1(\mathbf{x})\}$, and $\{\mathbf{x}_k, \mathbf{y}_k\} = \{\mathcal{T}_{k-1}(\mathbf{x}), \mathcal{T}_k(\mathbf{x})\}$, $k > 1$, where \mathcal{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 $\mathcal{T}(\mathcal{W}^T, \alpha^T)$, 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 observes two phenomenons that can be attrbute to the biased supervision, *i.e.* *candidate*

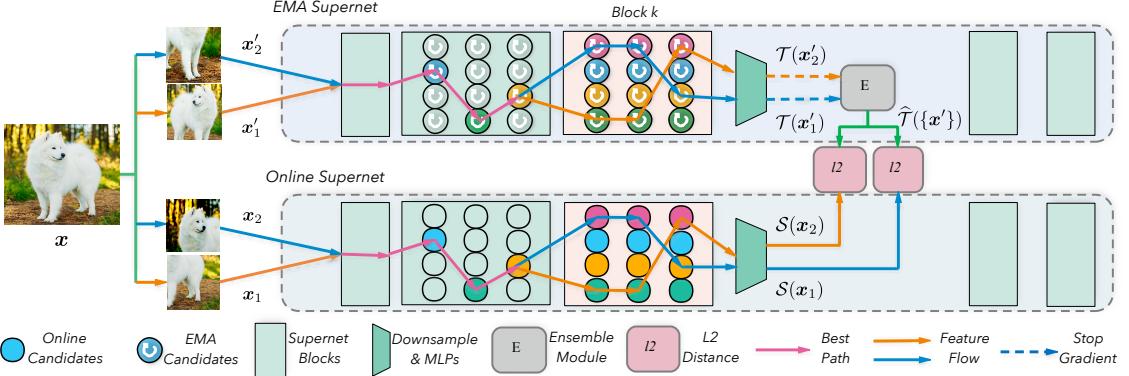


Figure 2: Illustration of the Siamese supernets training with ensemble bootstrapping.

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 $\{\mathcal{S}(\mathcal{W}, \mathcal{A}), \mathcal{T}(\mathcal{W}^T, \mathcal{A}^T)\}$, the first step to cast off the yoke of the teacher architecture is to assign $\alpha^T = \mathcal{A}$, 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 $\{\mathcal{S}(\mathcal{W}, \mathcal{A}; x_1), \mathcal{T}(\mathcal{W}^T, \mathcal{A}; x_2)\}$, 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 [10, 11] (*i.e.* $\mathcal{W}^T = \mathcal{W}$) or form a mean teacher scheme with Exponential Moving Average (EMA) [24, 21] (*i.e.* $\mathcal{W}^T = \mathcal{W}^*$, where $\mathcal{W}_t^* = \tau \mathcal{W}_{t-1}^* + (1 - \tau) \mathcal{W}_t$ represents the temporal average of \mathcal{W} , with t being a training timestamp, and τ denoting the momentum factor that controls the updating speed of \mathcal{W}^*). By learning representation from the mean teacher, analogous to the simple yet powerful BYOL [21], our supernet can be optimized in an unsupervised manner without relying on a fully supervised teacher network:

$$\mathcal{W}_k^* = \arg \min_{\mathcal{W}_k} \mathcal{L}_{\text{train}}(\{\mathcal{W}_k, \mathcal{W}_k^*\}, \mathcal{A}_k; x_k). \quad (2)$$

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 reception 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 [22] or fair path [15]. 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 hardship, leading to training in-

stability and inaccurate architecture ratings. To address this problem, we propose an unsupervised supernet training scheme, named *ensemble bootstrapping*.

Considering $|p|$ sub-networks $\{\alpha_p\} \subset \mathcal{A}_k$ sampled from the k -th block of the search space \mathcal{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 [59, 60] 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:

$$\widehat{\mathcal{T}}_k(\{\alpha_p\}; \{x'_p\}) = \frac{1}{|p|} \sum_{p=1}^{|p|} \mathcal{T}_k(\mathcal{W}^*, \alpha_p; x'_p). \quad (3)$$

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

$$\begin{aligned} \mathcal{W}_k^* &= \arg \min_{\mathcal{W}_k} \sum_{p=1}^{|p|} \mathcal{L}_{\text{train}}(\{\mathcal{W}_k, \mathcal{W}_k^*\}, \{\alpha_p\}; x), \\ \text{where } \mathcal{L}_{\text{train}}(\{\mathcal{W}_k, \mathcal{W}_k^*\}, \{\alpha_p\}; x) &= \left\| \mathcal{S}_k(\mathcal{W}_k, \alpha_p; x_p) - \widehat{\mathcal{T}}_k(\mathcal{W}_k^*, \{\alpha_p\}; x'_p) \right\|_2^2. \end{aligned} \quad (4)$$

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}_{val} for searching phase.

To evaluate the performance of a network trained with contrastive self-supervision, previous works [24, 10, 21, 11] 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, 87] 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

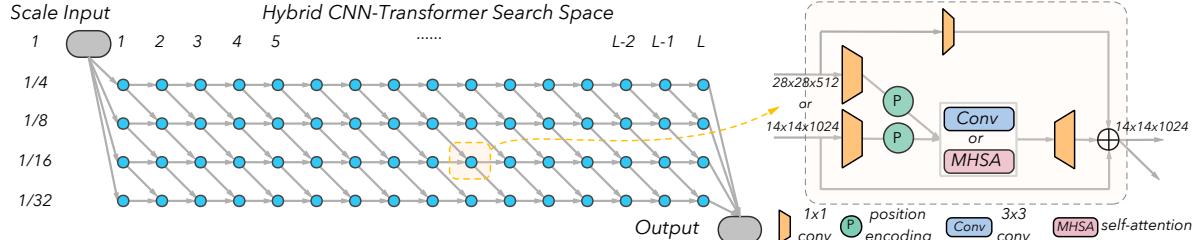


Figure 3: Illustration of the fabric-like Hybrid CNN-transformer Search Space with flexible down-sampling positions.

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* [12, 54], 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 α_p . Additionally, one pair of views $\{x_1, x_2\}$ for each validation samples 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:

$$\widehat{\mathcal{S}}_k(\{\alpha_p\}; x_2) = \frac{1}{|p|} \sum_{p=1}^{|p|} \mathcal{S}_k(\alpha_p; x_2). \quad (5)$$

In practice, by dividing the supernet 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 \mathcal{A}_k , and the whole searching process is finished in a single step:

$$\alpha^* = \arg \min_{\forall \alpha \in \mathcal{A}} \sum_{k=1}^{|k|} \lambda_k \mathcal{L}_{\text{val}}(\alpha; x_k) \quad (6)$$

where $\mathcal{L}_{\text{val}}(\alpha; x) = \left\| \mathcal{S}_k(\alpha; x_1) - \widehat{\mathcal{S}}_k(\mathcal{A}_k; x_2) \right\|_2^2$.

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 [25] as the CNN candidate building block. In parallel, we design a lightweight and robust transformer building block **ResAtt** based on the pluggable *BoTBlock* [62] and *NLBlock* [68].

Computation Balancing with Implicit Position Encodings. To facilitate fair and meaningful competition, can-

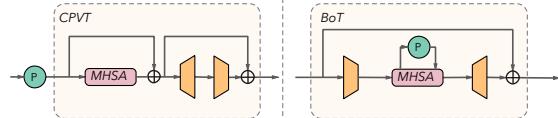


Figure 4: Transformer blocks in CPVT [14] and BoT [62].

dicate 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 [14], 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 the original $\mathcal{O}(CW^3)$ to $\mathcal{O}(CW^2)$, 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* [57] 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 [20, 66, 14], CNNs [25, 31] and hybrid CNN-transformers [62] at different scales.

Method	MAdds	Steptime	Top-1 (%)	Top-5 (%)
ResNet50 [25]	4.1B	100ms	77.7	93.9
ViT-B/32 [20]	-	68ms	73.4	-
ViT-B/16 [20]	17.6B	158ms	77.9	-
BoT50 [62]	4.0B	120ms	78.3	94.2
R50-T Conv-Only	4.1B	104ms	78.2	94.2
ViT-T/32 Att-Only	2.9B	92ms	74.5	91.7
ViT-T/16 Att-Only	3.2B	96ms	76.5	93.0
BoT50-T Hybrid	3.9B	103ms	79.5	94.8
Random-T Hybrid	3.7B	84ms	76.7	93.1
BossNet-T0 w/o SE	3.4B	101ms	80.5	95.0
SENet50 [25, 31]	4.1B	129ms	79.4	94.6
EffNetB1 [64]	0.7B	131ms	79.1	94.4
DeiT-S [66]	10.1B	84ms	79.8	-
BoT50 + SE [62]	4.0B	149ms	79.6	94.6
DNA-T [37]	3.9B	121ms	80.3	95.0
UnNAS-T [41]	3.7B	104ms	79.8	94.6
BossNet-T0	3.4B	115ms	80.8	95.2
BossNet-T0 ↑	5.7B	147ms	81.6	95.6
SENet101 [25, 31]	7.8B	218ms	81.4	95.7
EffNetB2 [64]	1.0B	143ms	80.1	94.9
ViT-L/16 [20]	63.6B	168ms	81.1	-
DeiT-B [66]	17.6B	152ms	81.8	-
BoTNet-S1-59 [62]	7.3B	184ms	81.7	95.8
T2T-ViT-19 [79]	8.9B	158ms	81.9	-
TNT-S [23]	5.2B	468ms	81.3	95.6
BossNet-T1	7.9B	156ms	82.2	95.8
BossNet-T1 ↑	10.5B	165ms	82.5	96.0

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. \uparrow : Directly tested on larger input size without finetuning (*i.e.* 288 for BossNet-T0 \uparrow and 256 for BossNet-T1 \uparrow).

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, 37] and NATS-Bench size search space S_S [17]. The datasets we use to evaluate and analyze our method are ImageNet [16], CIFAR-10 and CIFAR-100 [36]. We train each block of the supernet for 20 epochs, including one linear warm-up epoch. We randomly sample four paths in each training step. See Appendix A.2 for more implementation details.

5.1. Searching for Hybrid CNN-transformer

Analysis 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 [25, 20, 62], 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 \times** 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.

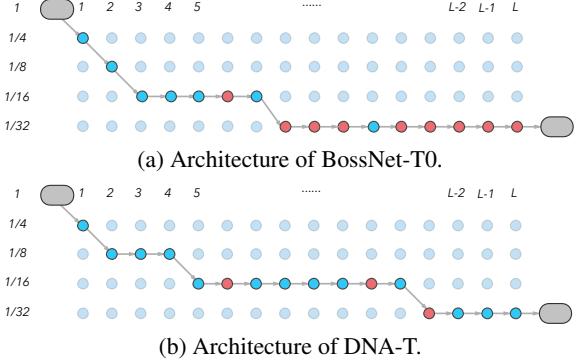


Figure 5: Visualization of architectures searched by Boss-NAS and DNA [37] in HyTra search space. Blue nodes denotes **ResConv** and red nodes denotes **ResAtt**.

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 \times** 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 \times** 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 [37].

Finally, when extended to larger model size and input size, the family of BossNet-T models maintain their superiority. By removing the downsampling in the last stage of BossNet-T0 (same scheme as BoTNet-S1 [62]), we have BossNet-T1, which achieves **82.2%** accuracy, surpassing EfficientNet-B2 by **2.1%**. By directly testing on larger input resolutions without finetuning, BossNet-T0 \uparrow (on 288 \times 288 input size) achieves **81.6%** top-1 accuracy, and outperforms BoTNet50 + SE by **2.0%** with similar runtime; BossNet-T1 \uparrow (on 256 \times 256 input size) achieves **82.5%** top-1 accuracy, surpassing T2T-ViT-19 and EfficientNet-B2 by **0.6%** and **2.4%** with comparable steptime, respectively.

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 architectural bias from the teacher supervision. Without using the teacher model, our method successfully avoids this bias.

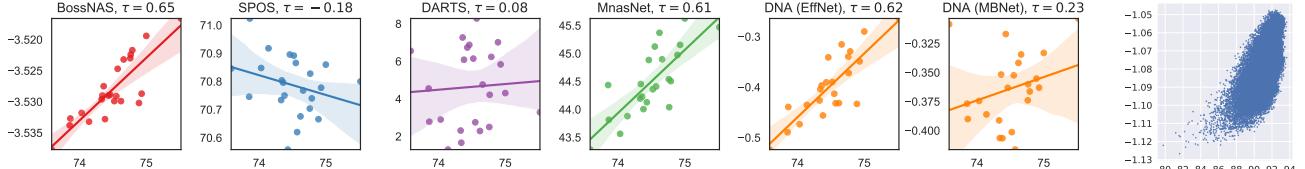


Figure 6: **Left:** Ranking correlations of 6 different NAS methods on **MBConv Search Space**. **Right:** Architecture ranking of BossNAS on **NATS-Bench S_S** . In all the diagrams, x-axis denotes ground truth accuracy; y-axis denotes evaluation metrics.

Method	MAdds (M)	Top-1 (%)	Top-5 (%)
FairNAS-A [15]	388M	75.3	92.4
ProxylessNAS [7]	465M	75.1	92.5
FBNet-C [70]	375M	74.9	-
SPOS [22]	472M	74.8	-
RLNAS [87]	473M	75.6	92.6
BossNet-M1 w/o SE	475M	76.2	93.0
MobileNetV3 [29]	219M	75.2	-
MnasNet-A3 [63]	403M	76.7	93.3
EfficientNet-B0 [64]	399M	76.3	93.2
DNA-b [37]	406M	77.5	93.3
BossNet-M2	403M	77.4	93.6

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

Method	Search Cost	τ	ρ	R
SPOS [22]	8.5 Gds	-0.18	-0.27	-0.29
DARTS [43]	50 Gds	0.08	0.14	0.06
MnasNet [63]	288 Tds	0.61	<u>0.77</u>	0.78
DNA [37] (EffNetB0)	8.5 Gds	<u>0.62</u>	<u>0.77</u>	<u>0.83</u>
DNA [37] (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**. (Gds: GPU days; Tds: TPU days)

5.2. Results on MBConv Search Space

To further prove the effectiveness and generalization ability of BossNAS, we compare it with a wide range of NAS methods on MBConv search space.

Performance of searched models. As shown in Tab. 2, our searched models, BossNet-M, achieve competitive results in search spaces with and without SE module. In the search space *without* SE, BossNet-M1, searched under constraint of 475M MAdds, outperforms SPOS [22] and another recent unsupervised NAS method, RLNAS [87] by **1.4%** and **0.6%**, respectively. In the search space *with* SE, BossNet-M2, under constraint of 405M MAdds, outperforms the popular EfficientNet [64] by **1.1%**, and is also competitive with our supervised counterpart, DNA [37]. Note that candidate building blocks in MBConv search space are quite similar, concealing the *candidate preference* phenomenon in [37].

Architecture rating accuracy. 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 [37] to calculate the architecture rating accuracy, *i.e.* the ranking correlation between the

Method	C-10	C-100	τ	ρ	R
FBNet v2 [67]	93.14	70.72	-	-	-
TuNAS [5]	92.78	70.11	-	-	-
CE [27]	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 S_S** (C-10: **CIFAR-10**, C-100: **CIFAR-100**).

predicted architecture ranking and the ground truth model ranking. We use three different ranking correlation metrics: Kendall Tau (τ) [34], 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 obtains the highest rating accuracies with **0.65 τ** among *sota* NAS methods, while addressing two problems.

First, classic weight sharing methods, SPOS [22] and DARTS [43], fails to achieve reasonable ranking correlation despite their lower search costs, while the multi-trial method, MnasNet [63], 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 \times** acceleration.

Second, supervised block-wise NAS method, DNA [37], fails to achieve high rating accuracies when using a teacher largely different from the candidates (MobileNetV1 [30] vs. EfficientNet-based candidates [64]), 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 S_S

For NATS-Bench size search space S_S , experiments are conducted on two datasets: CIFAR-10 and CIFAR-100. Candidates of different channel numbers in our supernet share the weights in a slimmable manner [78, 77, 76, 38, 9].

Performance of searched models. After searching on our supernet, we look up the performance of searched models in NATS-Bench S_S for fair comparison. The results are shown in Tab. 4. Our BossNAS outperforms recent NAS methods [67, 5] designed particularly for network size search spaces, proving the generalization ability of our method on specified search spaces and relatively small datasets.

Architecture rating accuracy. We rate all the 32768 architectures in the search space to compare with their ground

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**.

truth accuracies in the benchmark on CIFAR-10 dataset. As shown in Fig. 6 right, all the architectures in the search space forms a dense, spindle-shaped pattern, proving the 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 [27], which is trained with ground truth architecture accuracies, by a large gap (*i.e.* **0.16 τ** and **0.19 R**). More analysis on NATS-Bench \mathcal{S}_S could be found in Appendix A.3.

5.4. Ablation Study

In this section, we perform extensive ablation studies on MBConv search space and ImageNet to analyze our proposed training and evaluation methods separately.

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 [37] (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 the 3-*rd* and 5-*th* line, we can see that replacing our proposed Unsupv. EB with the naive Unsupv. bootstrap. scheme, the architecture rating accuracy drops sharply by **0.53 τ** . 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 also 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 the 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

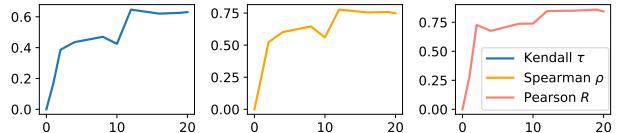


Figure 7: Ranking correlations during supernet training.

weight sharing linear classifier to evaluate each architecture. (4) Our unsupervised evaluation metric (Unsupv. eval) rate architectures by its distance to the ensemble probability center of the whole search 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 τ**).

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 \mathcal{S}_S .

6. Conclusion

In this work, we present BossNAS, a general, unsupervised NAS method with the *ensemble bootstrapping* training technique and an *unsupervised evaluation metric*. 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 82.5% accuracy on ImageNet, surpassing EfficientNet by 2.4% with comparable compute time.

Acknowledgement

This work was supported in part by National Key R&D Program of China under Grant No. 2020AAA0109700 and the funding of “Leading Innovation Team of the Zhejiang Province” (2018R01017). Dr Xiaojun Chang is partially supported by Australian Research Council (ARC) Discovery Early Career Research Award (DECRA) under grant no. DE190100626 and Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC).

References

- [1] Youhei Akimoto, Shinichi Shirakawa, Nozomu Yoshinari, Kento Uchida, Shota Saito, and Kouhei Nishida. Adaptive stochastic natural gradient method for one-shot neural architecture search. In *ICML*, 2019. 12
- [2] Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. In *ICLR*, 2017. 2, 12
- [3] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention. In *ICLR*, 2021. 1
- [4] Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, and Quoc V. Le. Understanding and simplifying one-shot architecture search. In *ICML*, 2018. 2, 12
- [5] Gabriel Bender, Hanxiao Liu, Bo Chen, Grace Chu, Shuyang Cheng, Pieter-Jan Kindermans, and Quoc V. Le. Can weight sharing outperform random architecture search? an investigation with tunas. In *CVPR*, 2020. 7
- [6] Andrew Brock, Theodore Lim, James M. Ritchie, and Nick Weston. SMASH: one-shot model architecture search through hypernetworks. In *ICLR*, 2018. 2, 12
- [7] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. In *ICLR*, 2019. 3, 6, 7, 12
- [8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1
- [9] Minghao Chen, Houwen Peng, Jianlong Fu, and Haibin Ling. AutoFormer: Searching transformers for visual recognition. In *ICCV*, 2021. 1, 7, 13
- [10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3, 4
- [11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 3, 4
- [12] Yukang Chen, Gaofeng Meng, Qian Zhang, Shiming Xiang, Chang Huang, Lisen Mu, and Xinggang Wang. RENAS: reinforced evolutionary neural architecture search. In *CVPR*, 2019. 2, 5
- [13] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. In *NeurIPS*, 2020. 12
- [14] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv:2102.10882*, 2021. 1, 5
- [15] Xiangxiang Chu, Bo Zhang, and Ruijun Xu. FairNAS: Rethinking evaluation fairness of weight sharing neural architecture search. In *ICCV*, 2021. 3, 4, 7, 12
- [16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6, 13
- [17] Xuanyi Dong, Lu Liu, Katarzyna Musial, and Bogdan Gabrys. NATS-Bench: Benchmarking nas algorithms for architecture topology and size. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. doi:[10.1109/TPAMI.2021.3054824](https://doi.org/10.1109/TPAMI.2021.3054824). 2, 3, 6, 12, 13
- [18] Xuanyi Dong and Yi Yang. Searching for a robust neural architecture in four GPU hours. In *CVPR*, 2019. 12
- [19] Xuanyi Dong and Yi Yang. NAS-Bench-201: Extending the scope of reproducible neural architecture search. In *ICLR*, 2020. 3
- [20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 5, 6
- [21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. 3, 4, 13
- [22] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. In *ECCV*, 2020. 4, 7, 12
- [23] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv:2103.00112*, 2021. 6
- [24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3, 4
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 5, 6, 13
- [26] Shuteng He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based object re-identification. In *ICCV*, 2021. 1
- [27] Daniel Hesslow and Iacopo Poli. Contrastive embeddings for neural architectures. *arXiv:2102.04208*, 2021. 3, 7, 8, 13
- [28] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 3
- [29] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *ICCV*, 2019. 2, 3, 7, 12
- [30] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 2017. 1, 7
- [31] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 1, 5, 6
- [32] Yiming Hu, Yuding Liang, Zichao Guo, Ruosi Wan, Xiangyu Zhang, Yichen Wei, Qingyi Gu, and Jian Sun. Angle-based search space shrinking for neural architecture search. In *ECCV*, 2020. 3

- [33] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv:2102.07074*, 2021. 1
- [34] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 7
- [35] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 2, 13
- [36] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master’s thesis, Department of Computer Science, University of Toronto*, 2009. 6, 13
- [37] Changlin Li, Jiefeng Peng, Liuchun Yuan, Guangrun Wang, Xiaodan Liang, Liang Lin, and Xiaojun Chang. Blockwisely supervised neural architecture search with knowledge distillation. In *CVPR*, 2020. 2, 3, 6, 7, 8, 12, 13
- [38] Changlin Li, Guangrun Wang, Bing Wang, Xiaodan Liang, Zhihui Li, and Xiaojun Chang. Dynamic Slimmable Network. In *CVPR*, 2021. 7, 13
- [39] Xiang Li, Chen Lin, Chuming Li, Ming Sun, Wei Wu, Junjie Yan, and Wanli Ouyang. Improving one-shot nas by suppressing the posterior fading. In *CVPR*, 2020. 2, 3
- [40] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *CVPR*, 2019. 3
- [41] Chenxi Liu, Piotr Dollár, Kaiming He, Ross Girshick, Alan Yuille, and Saining Xie. Are labels necessary for neural architecture search? In *ECCV*, 2020. 2, 4, 6, 13
- [42] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *ECCV*, 2018. 3, 12
- [43] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architecture search. In *ICLR*, 2019. 2, 3, 7, 12
- [44] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 13
- [45] Hartmut Maennel, Ibrahim M Alabdulmohsin, Ilya O Tolstikhin, Robert Balduck, Olivier Bousquet, Sylvain Gelly, and Daniel Keysers. What do neural networks learn when trained with random labels? In *NeurIPS*, 2020. 3
- [46] Bert Moons, Parham Noorzarad, Andrii Skliar, Giovanni Mariani, Dushyant Mehta, Chris Lott, and Tijmen Blankevoort. Distilling optimal neural networks: Rapid search in diverse spaces. *arXiv:2012.08859*, 2020. 2, 3, 12
- [47] Renato Negrinho and Geoffrey J. Gordon. Deeparchitect: Automatically designing and training deep architectures. *arXiv:1704.08792*, 2017. 2
- [48] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2
- [49] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 3
- [50] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018. 1
- [51] Jiefeng Peng, Jiqi Zhang, Changlin Li, Guangrun Wang, Xiaodan Liang, and Liang Lin. Pi-NAS: Improving neural architecture search by reducing supernet training consistency shift. In *ICCV*, 2021. 3, 12
- [52] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *ICML*, 2018. 3, 12
- [53] Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. In *ICML*, 2018. 2
- [54] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Regularized evolution for image classifier architecture search. In *AAAI*, 2019. 3, 5, 12
- [55] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys*, 2021. 12
- [56] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 3
- [57] Shreyas Saxena and Jakob Verbeek. Convolutional neural fabrics. In *NeurIPS*, 2016. 3, 5
- [58] Christian Sciuto, Kaicheng Yu, Martin Jaggi, Claudiu Musat, and Mathieu Salzmann. Evaluating the search phase of neural architecture search. In *ICLR*, 2020. 3
- [59] Zhiqiang Shen, Zhankui He, and Xiangyang Xue. Meal: Multi-model ensemble via adversarial learning. In *AAAI*, 2019. 4
- [60] Zhiqiang Shen and Marios Savvides. Meal v2: Boosting vanilla resnet-50 to 80%+ top-1 accuracy on imagenet without tricks. *arXiv:2009.08453*, 2020. 4
- [61] Han Shi, Renjie Pi, Hang Xu, Zhenguo Li, James Kwok, and Tong Zhang. Bridging the gap between sample-based and one-shot neural architecture search with bonas. In *NeurIPS*, 2020. 12
- [62] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *CVPR*, 2021. 1, 2, 5, 6
- [63] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile. In *CVPR*, 2019. 2, 3, 7, 12
- [64] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 1, 2, 3, 6, 7, 12, 13
- [65] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv:1906.05849*, 2019. 3
- [66] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1, 5, 6, 13
- [67] Alvin Wan, Xiaoliang Dai, Peizhao Zhang, Zijian He, Yuan-dong Tian, Saining Xie, Bichen Wu, Matthew Yu, Tao Xu, Kan Chen, et al. Fbnetv2: Differentiable neural architecture search for spatial and channel dimensions. In *CVPR*, 2020. 7
- [68] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 5

- [69] Chen Wei, Yiping Tang, Chuang Niu, Haihong Hu, Yue Wang, and Jimin Liang. Self-supervised representation learning for evolutionary neural architecture search. *arXiv:2011.00186*, 2020. 3
- [70] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *CVPR*, 2019. 3, 7, 12
- [71] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3
- [72] Shen Yan, Yu Zheng, Wei Ao, Xiao Zeng, and Mi Zhang. Does unsupervised architecture representation learning help neural architecture search? In *NeurIPS*, 2020. 3
- [73] Antoine Yang, Pedro M. Esperança, and Fabio Maria Carlucci. NAS evaluation is frustratingly hard. In *ICLR*, 2020. 3
- [74] Chris Ying, Aaron Klein, Eric Christiansen, Esteban Real, Kevin Murphy, and Frank Hutter. Nas-bench-101: Towards reproducible neural architecture search. In *ICML*, 2019. 3
- [75] Yang You, Igor Gitman, and Boris Ginsburg. Scaling sgd batch size to 32k for imagenet training. *arXiv:1708.03888*, 2017. 13
- [76] Jiahui Yu and Thomas Huang. Autoslim: Towards one-shot architecture search for channel numbers. *arXiv:1903.11728*, 2019. 7, 13
- [77] Jiahui Yu and Thomas S. Huang. Universally slimmable networks and improved training techniques. In *ICCV*, 2019. 7, 13
- [78] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas S. Huang. Slimmable neural networks. In *ICLR*, 2019. 7, 13
- [79] 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. In *ICCV*, 2021. 1, 6, 13
- [80] Arber Zela, Julien Siems, and Frank Hutter. Nas-bench-1shot1: Benchmarking and dissecting one-shot neural architecture search. In *ICLR*, 2019. 3, 12
- [81] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530*, 2016. 3
- [82] Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, Zongyuan Ge, and Steven W. Su. Differentiable neural architecture search in equivalent space with exploration enhancement. In *NeurIPS*, 2020. 12
- [83] 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. 12
- [84] Mingyang Zhang and Linlin Ou. Stage-wise channel pruning for model compression. *arXiv:2011.04908*, 2020. 2, 3
- [85] 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. 2, 3
- [86] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2
- [87] Xuanyang Zhang, Pengfei Hou, Xiangyu Zhang, and Jian Sun. Neural architecture search with random labels. In *CVPR*, 2021. 2, 3, 4, 7, 12
- [88] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian 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. In *CVPR*, 2021. 1
- [89] Zhao Zhong, Junjie Yan, Wei Wu, Jing Shao, and Cheng-Lin Liu. Practical block-wise neural network architecture generation. In *CVPR*, 2018. 2
- [90] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 1
- [91] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *ICCV*, 2019. 3
- [92] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In *ICLR*, 2017. 2, 12
- [93] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *CVPR*, 2018. 3

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

Supplementary Material

Changlin Li¹, Tao Tang², Guangrun Wang^{3,4}, Jiefeng Peng³, Bing Wang⁵,
Xiaodan Liang^{2*}, Xiaojun Chang⁶

¹GORSE Lab, Dept. of DSAI, Monash University ²Sun Yat-sen University ³DarkMatter AI Research

⁴University of Oxford ⁵Alibaba Group ⁶RMIT University

changlin.li@monash.edu,

{trent.tangtao,wanggrun,jiefengpeng,xdliang328}@gmail.com,
fengquan.wb@alibaba-inc.com, xiaojun.chang@rmit.edu.au

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 [92, 2, 54, 63, 42, 83] 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) [92, 2, 54] or by training for the first few epochs (*e.g.* 5 epochs) [63] on ImageNet. To avoid repeated training of candidate networks, **weight-sharing NAS** methods [7, 43, 18, 1, 6, 13, 82] optimize a *supernet* that encodes the whole search space, then rate each candidate architecture according to its weights inherited from the supernet. Among them, *gradient-based* approaches [43, 7, 70] and *sampler-based* approaches [52, 61] jointly optimize the weight of the supernet and the factors (or agent) used to choose the architecture; for their part, *one-shot* approaches¹ [22, 15, 6, 4, 51] optimize the supernet before performing a search with the frozen supernet weights. We refer to [55] 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 2.8×10^6 .

- **MBCov search space.** MobileNet-like search space and its variations are generally used as benchmarks for recent NAS methods [63, 29, 64, 7, 70, 15, 37, 46, 87]. Following Li *et. al.* [37], 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 \mathcal{S}_S .** The NATS-Bench *size* search space \mathcal{S}_S [17] is a channel configuration search space built

*Corresponding Author.

¹In this paper, following the pioneering works SMASH [6] and One-shot [4], when we refer to one-shot NAS methods, we are discussing those incorporating two-stage (*i.e.*, a supernet training stage and a searching stage) weight-sharing methods rather than the general weight-sharing NAS discussed in [80].

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 $\{8, 16, 24, 32, 40, 48, 56, 64\}$. \mathcal{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 [78, 77, 76, 38, 9]. We divide the supernet into 3 blocks, according to spatial size.

Datasets. The datasets we use to evaluate and analyze our method include ImageNet [16], CIFAR-10 and CIFAR-100 [36]. **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** [36], we use the splits proposed in NATS-Bench [17]. CIFAR-10 is divided into 25 K train set, 25 K val set, and 10 K test set. CIFAR-100 is devided 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 \mathcal{S}_S [17].

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 [21], we use the LARS optimizer [75] 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 similar with DeiT [66], as we found it robust for both CNNs and transformers. More specifically, we use AdamW optimizer with 1e-3 initial learning rate and cosine learning rate scheduler, for a total batch size of 1024. Weight decay is set to 0.05. We use model EMA with decay rate 0.99996 following [79]. Please refer to DeiT [66] for more details on data-augmentation and regularization.

For ImageNet retraining of **BossNet-M models**, we follow closely to EfficientNet [64]. 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 [64] for more details of other settings.

Re-implementation of other NAS methods on HyTra.

For DNA [37], we use ResNet-50 [25] 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

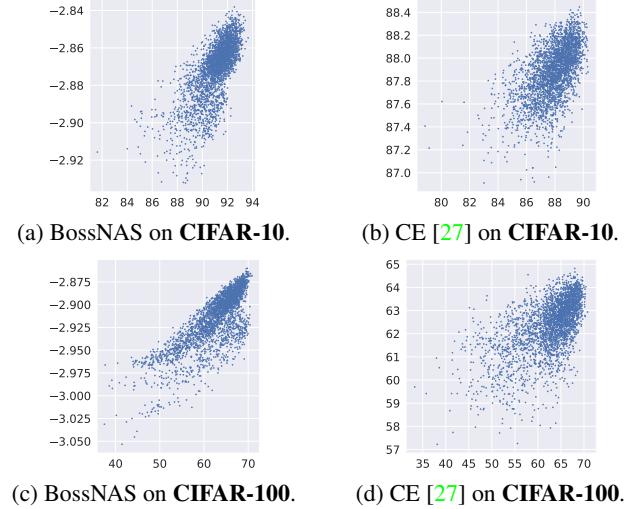


Figure 8: Comparison of architecture rating and its true accuracy of our BossNAS and CE [27] on **NATS-Bench \mathcal{S}_S** with **CIFAR datasets**.

Dataset	Method	τ	ρ	R
CIFAR-10	CE [27]	0.42	0.60	0.59
	BossNAS	0.53	0.73	0.72
CIFAR-100	CE [27]	0.43	0.60	0.60
	BossNAS	0.59	0.76	0.79

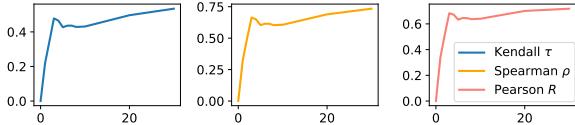
Table 6: Architecture rating accuracy on **NATS-Bench \mathcal{S}_S** with **CIFAR datasets**.

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 [37].

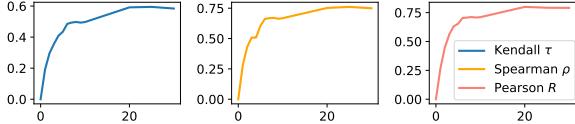
For UnNAS [41], we adopt *rotation prediction* [35] (Rot) pretext task, for its simplicity. Following [41], we use three extra stride-2 convolution layers at the beginning of the supernet to reduce spatial resolution. The supernet is trained for 2 epochs as in [41].

A.3. Additional Analysis on NATS-Bench \mathcal{S}_S

Architecture rating comparison. We compare with the predictor-based NAS method CE [27] 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 \mathcal{S}_S [17]. 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 τ**) in both datasets.

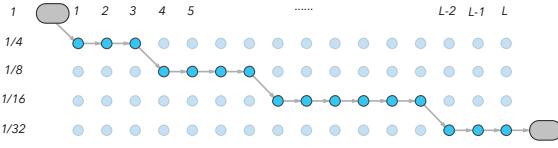


(a) Ranking correlations during supernet training on **CIFAR-10**.

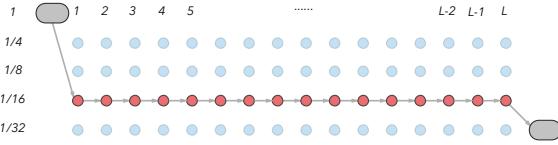


(b) Ranking correlations during supernet training with **CIFAR-100**.

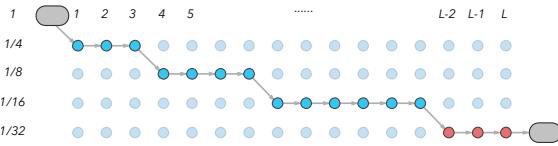
Figure 9: Convergence behavior of BossNAS on NATS-Bench $\mathcal{S}_{\mathcal{S}}$ and CIFAR datasets.



(a) Architecture of ResNet50-T.



(b) Architecture of ViT-T/16.



(c) Architecture of BoTNet-T.

Figure 10: Visualization of Human-designed Architectures in HyTra. Blue nodes denotes `ResConv` and red nodes denotes `ResAtt`.

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 ρ** , smoothly and quickly within only 20 epochs on CIFAR-100, and continues to be stable for the subsequent 10 epochs.

A.4. Visualization of Human-designed Architectures in HyTra

The architectures of ResNet50-T, ViT-T/16 and BoTNet50-T from our HyTra search space are illustrated in Fig. 10. Their architectures follow as closely as possible to the architectures of their prototypes.