DisWOT: Student Architecture Search for Distillation WithOut Training

Peijie Dong[1†] Lujun Li[2†*] Zimian Wei[1†]
[1] National University of Defense Technology, [2] Chinese Academy of Sciences
[1]{dongpeijie, weizimian16}@nudt.edu.cn, [2]lilujunai@gmail.com

Abstract

Knowledge distillation (KD) is an effective training strategy to improve the lightweight student models under the guidance of cumbersome teachers. However, the large architecture difference across the teacher-student pairs limits the distillation gains. In contrast to previous adaptive distillation methods to reduce the teacher-student gap, we explore a novel training-free framework to search for the best student architectures for a given teacher. Our work first empirically show that the optimal model under vanilla training cannot be the winner in distillation. Secondly, we find that the similarity of feature semantics and sample relations between random-initialized teacher-student networks have good correlations with final distillation performances. Thus, we efficiently measure similarity matrixs conditioned on the semantic activation maps to select the optimal student via an evolutionary algorithm without any training. In this way, our student architecture search for Distillation WithOut Training (DisWOT) significantly improves the performance of the model in the distillation stage with at least 180× training acceleration. Additionally, we extend similarity metrics in DisWOT as new distillers and KD-based zero-proxies. Our experiments on CIFAR, ImageNet and NAS-Bench-201 demonstrate that our technique achieves state-of-the-art results on different search spaces. Our project and code are available at https://lilujunai.github.io/DisWOT-CVPR2023/.

1. Introduction

Despite the remarkable achievements of Deep Neural Networks (DNNs) in numerous visual recognition tasks [52, 64–68], they usually lead to heavy costs of memory, computation, and power at model inference due to their large numbers of parameters. To address this issue, Knowledge Distillation (KD) has been proposed as a means of transferring knowledge from a high-capacity teacher model to a low-capacity target student model, providing a more optimal accuracy-efficiency trade-off during runtime [4, 7, 76]. The original KD method [22] utilizes the logit outputs of the teacher network as the source of knowledge. Subsequent studies [2, 21, 24, 26, 54, 61, 73] have focused on extracting informative knowledge based on intermediate feature representations. However, as the gap in capacity between students and teachers increases, existing KD methods are unable to improve results, particularly in tasks that depend on large-scale visual models such as VIT and GPT-3 [5, 16]. For example, as shown in Figure 2 (Left), the large teacher (e.g., ResNet110) lead to worse performance for the fixed student than the relatively smaller one (e.g., ResNet56).

To solve this issue, adaptive KD methods have been proposed in terms of training paradigms (e.g., early stop [10]) and architectural adaptations (e.g., assistant teacher [43] and architecture search [40]), respectively. However, they are ineffective in improving distillation performance or in-

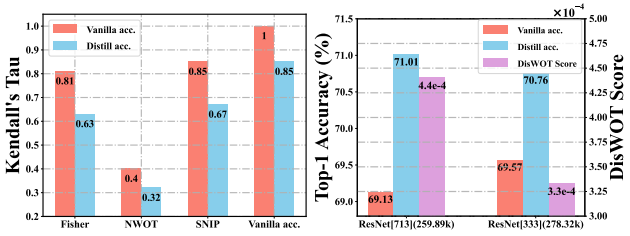

Figure 1. Left: Ranking correlation of proxies in zero-cost NAS with vanilla and distillation accuracy. Right: Vanilla accuracy, distillation accuracy, prediction scores of DisWOT for ResNet[7,1,3] and ResNet[3,3,3] on search space S_0.

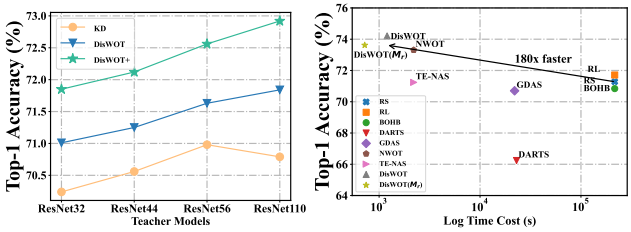

Figure 2. Left: KD [22], DisWOT, DisWOT† results for ResNet20 under different teachers. Right: Comparison of distill accuracy & training efficiency with other NAS methods on NAS-Bench-201.

*Corresponding author, † equal contribution, PD conducted main experiments, LL proposed ideas and led the project & writing.

volve enormous training costs in the additional model training and search process. In sharp contrast to these methods, we tackle this challenging problem from a new perspective regarding training-free architecture search. To achieve this goal, we construct a search space S_0 for ResNet-like models with different depth configurations and obtain vanilla and distill performance for each candidate in S_0 by individual training. Then, we evaluate the ranking correlation between predicted scores of training-free search methods and the actual performance of each student model. Surprisingly, as shown in Figure 1 (Left), there are common ranking correlation loss ($10\% \downarrow\sim 20\% \downarrow$) for these methods in predicting distillation accuracy than vanilla accuracy. To clarify this, we carefully analyze the disparities in vanilla and distillation performance for each model: (1) for overall search space, vanilla accuracy only preserves 85% correlations with actual distillation performance. (2) for a particular instance, as shown in Figure 1 (Right), ResNet20 with 3 res-blocks in each stage (i.e., ResNet[3,3,3]) has more parameters and better standalone performance but is weaker than ResNet[7,1,3] in the distillation process. Considering that ResNet[7,1,3] has more layers than ResNet20, we seek to understand the above phenomenon regarding the vanilla-distillation accuracy gap from the perspective of semantic matching [37]. ResNet[7,1,3] enjoys a larger effective receptive field and more excellent matched knowledge with teacher, resulting in significant distillation gains. Encouraged by this understanding, we strive to design a new zero-proxy regarding the semantic matching of teacher-student. As a result, we find that the similarity scores of feature semantics and sample relations can outperform conventional zero-cost NAS in predicting final distillation accuracy (see the comparison of ranking correlation on search space S_0 in Table 8). As shown in Figure 1(Right), similarity scores are also consistent with distillation performance.

Drawing on the aforementioned observations, we introduce DisWOT, a simple yet effective training-free framework that finds the best student architectures for distilling the given teacher model. For better semantic matching in distillation, DisWOT leverages novel zero-cost metrics regarding the feature semantics and sample relations to select better student model. For the feature semantic similarity metric, we remark that randomly initialized models can localize objects well [6] and generate localization heatmaps via Grad-CAM [56] as reliable semantic information. Then, we measure the channel-wise similarity matrix of localization heatmaps and take the L_2 distance of the similarity matrix for the teacher-student model as the metric. For input samples, different models have diverse abilities to discriminate their relationships. To improve relational knowledge matching ability, we use the L_2 distance of sample-relation correlation matrix as a relation similarity metric. Finally, we search for student architectures using an evolutionary algorithm with semantic and relations similarity metrics. Then, the distillation process is implemented between the searched student and the pre-defined teacher. In addition, we leverage these metrics directly as new distillers to enhance the student, as the DisWOT†. Equipped with our train-free search and distillation design, our DisWOT and DisWOT† framework significantly improve the model's accuracy-latency tradeoff in inference with at least 180× training acceleration.

In principle, our DisWOT use higher-order statistics of teacher-student models to optimize the student architecture to fit a given teacher model. Its merits can be highlighted in three aspects: (1) In contrast to training-based student architecture search requires the individual or weight-sharing training, our DisWOT does not require the training of student models in the search phase. In addition, DisWOT is efficient to compute and easy to implement as it uses only the mini-batch data at initialization. (2) DisWOT is a teacher-aware search for distillation, which has better predictive distill accuracy than conventional NAS. (3) DisWOT exploits the distance of higher-order knowledge between the neural networks, bridging knowledge distillation and zero-proxy NAS. We further demonstrate the competitive ranking correlation of DisWOT among 10 knowledge distances in KD as zero-proxy for predicting vanilla accuracy in NAS-Bench-201. We anticipate that our work on KD-based zero-proxy can offer some assistance in furthering research endeavors related to KD and NAS.

We conduct extensive experiments on CIFAR-100, ImageNet, and the NAS-Bench-201 [14] dataset, demonstrating the superiority of our proposed approach. In contrast to experiments in traditional architectural search, we focus on final distillation accuracy instead of the vanilla accuracy for the student. The results show that our DisWOT can achieve better accuracy than traditional Zero-shot NAS in the same search space. Besides, by switching to a larger space, our DisWOT can obtain new state-of-the-art architectures. For example, in the same ResNet-like search space, we significantly improved 1.62% Top-1 accuracy over KD for ResNet50-ResNet18 pair under the same training settings. We also conducted comprehensive ablation studies to investigate how our method can use the predictability of zero-cost metrics to boost the distillation performance.

Main Contributions:

- By analyzing and exploring the discrepancy between teacher-student capability, we empirically show that their semantic similarities have a stronger correlation with the final distillation accuracy. This motivates us to propose a new student architecture search for the Distillation without Training (DisWOT) framework to reduce the teacher-student capability gap, which, to the best of our knowledge, is not achieved in the area of knowledge distillation.

- DisWOT proposes novel zero-cost metrics on similarity of feature semantics and sample relations and ensemble these metrics to select the optimal student via an evolutionary algorithm at the initial time. In the distillation stage, DisWOT achieves state-of-the-art performances in multiple datasets and search spaces.
- We further expand 10 kinds of knowledge distances including DisWOT as new universal KD-based zero proxies, which enjoy competitive predictive power with actual performance of models. We hope that our contributions in this endeavor may aid to some degree in advancing future research on KD and NAS.

2. Related Work and Background

In this section, we summarize existing knowledge distillation and architecture search methods and clarify their differences to our method.

2.1. General Formulation of Knowledge Distillation

The fundamental concept underlying Knowledge Distillation (KD) involves utilizing acquired knowledge (e.g., logits [31], feature values [30, 32, 33, 35, 70], and sample relations [45, 59]) from a high-capacity teacher to guide the training of a student model. The training dataset (X, Y) comprises training samples $X = x_i{}_{i=1}^{n}$ and their corresponding labels $Y = y_i{}_{i=1}^{n}$. Let f_T be the output logits of the fixed teacher T and let f_S be the output of student S, respectively. In KD, the student network f_S is trained by minimizing:

$$\mathcal{L}_S = \mathcal{L}_{CE}(f_S, Y) + \mathcal{L}_{KL}\left(f_S, f_T\right) + \mathcal{D}_f\big(\phi_S(x), \phi_T(x)\big), \quad (1)$$

where $\mathcal{L}_{CE}$ is the regular cross-entropy loss. $\mathcal{L}_{KL}$ represents Kullback-Leibler (KL) divergence. $\mathcal{D}_f(\cdot,\cdot)$ is the distance function measuring the difference of intermediate feature representations (see Table 1 for particular distillers).

Table 1. Comparison of recent distillers.

Method	Knowledge	Distance $\mathcal{D}_f(\cdot,\cdot)$
FitNets [54]	Feature representation	$\mathcal{L}_2$
AT [73]	Attention maps	$\mathcal{L}_2$
CC [49]	Instance relation	$\mathcal{L}_2$
NST [24]	Neuron selectivity patterns	$\mathcal{L}_{MMD}$
PKT [46]	Similarity probability distribution	$\mathcal{L}_{KL}$

Comparison with Other Adaptive KDs for Distillation Gap. DisWOT is the first train-free architecture search solution to reduce the teacher-student gap. Unlike training manners [10] and KD-loss designs , DisWOT utilizes the classic KD training configurations and distillers. In addition, DisWOT is free from the assistant teacher in ATKD [43], which involves a complex training routine and budget. AKD [40] searches student via reinforcement learning based on feedback from individual training of lots of models. As a completely alternative technical route to these training-based NAS [18], our training-free DisWOT builds on new zero-proxy and achieves $180\times \sim 1000\times$ training acceleration, which greatly improves its easy-to-use and flexibility.

Table 2. Formulation of NAS methods. $\mathcal{A}$ is the search space. A candidate architecture in the search space is denoted as $\alpha \in \mathcal{A}$, which corresponds to a neural architecture $S(\alpha, w)$ with weight w. $\mathcal{W}$ is the weight of the supernet. $_{train}$ and $_{val}$ are the loss functions on the training and validation sets, respectively.

Type	Evaluation	Formula
Training-based	Multi-trial Training	$\alpha^* = \arg\min_{\alpha\in\mathcal{A}} {}_{val}(S(\alpha, w_\alpha)),$ s.t. $w_\alpha = \arg\min_{w} {}_{train}(S(\alpha, w))$
	Weight sharing	$\alpha^* = \arg\min_{\alpha\in\mathcal{A}} {}_{val}(S(\alpha, \mathcal{W}_\mathcal{A}(\alpha))),$ s.t. $\mathcal{W}_\mathcal{A} = \arg\min_{\mathcal{W}} {}_{train}(S(\mathcal{A}, \mathcal{W}))$
Training-free	Zero-cost Proxy	$\alpha^* = \underset{\alpha\in\mathcal{A}}{ZeroProxy}(S(\alpha, w))$

Table 3. Comparison with different training-free NAS.

Type	Method	Teacher-aware	Objective
Prune-based	SNIP [28], Fisher [1], Synflow [58]	✗	Vanilla acc.
Activation-based	NWOT [42], Zen-NAS [36]	✗	Vanilla acc.
KD-based	DisWOT (ours)	✓	Distill acc.

2.2. Revisiting Architecture Search Methods

Neural Architecture Search (NAS) is emerged to reduce human efforts in architecture design and automate the discovery of high-performance networks. As formalized in Tab. 2, Multi-trial NAS methods [40, 79] train a large number of candidates individually, which leads to extensive resource consumption. To alleviate this, many NAS [8, 12, 23, 51] methods adopt a weight-sharing strategy within a single supernet to facilitate the simultaneous training of candidates. The supernet is trained for hundred of epochs by path sampling [11, 19] or compound optimization with architecture representations [39, 69]. As an orthogonal direction, zero-cost NAS methods [42, 71] focus on identifying well-performed architectures with training-free metrics. For example, NWOT [42] calculates the architecture score based on the kernel matrix of binary activations between small batches of samples.

Comparison with NAS with Teacher. Some training-based NAS [29, 50, 77] employ a teacher model to supervise supernet training to improve predictive ability in the search stage. However, these methods aim to improve vanilla accuracy, not for distillation, and they do not use the teacher model in the full training stage. In addition, without any

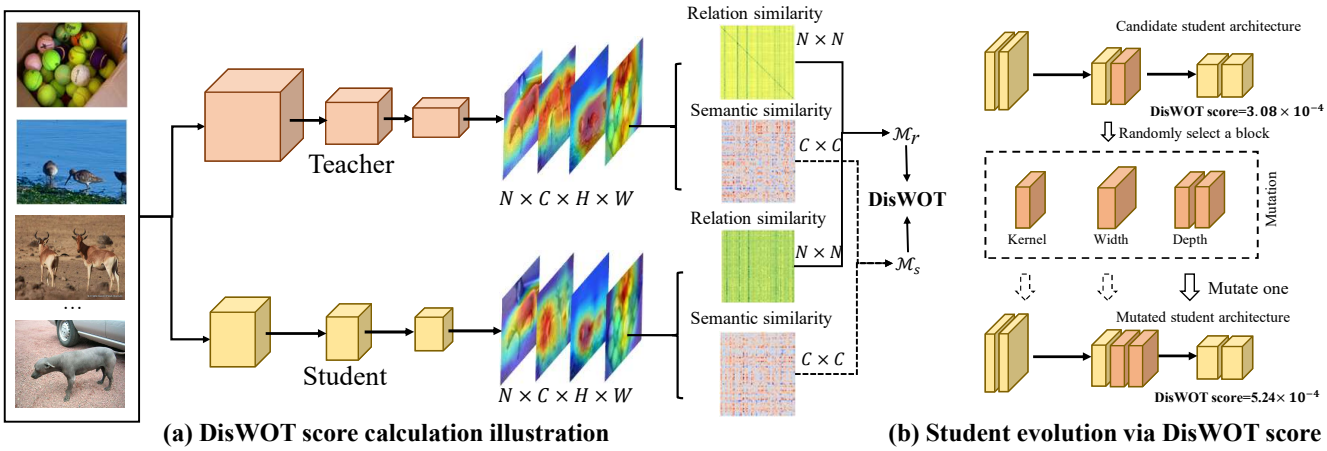

Figure 3. A schematic overview of our DisWOT, including (a) detailed calculation of the DisWOT scores and (b) evolution of the student architecture via the DisWOT scores. In search phase, DisWOT use semantic similarity metrics and relations similarity metrics to select good student for a given teacher. The semantic similarity metric is measured by l_2 distance of the channel-wise correlation matrix for Grad-cam activation maps. Similarly, the relation similarity matrix statistics the sample-wise correlation matrix distance of the randomly initialized teacher-student pairs. With the feedback from these metrics, the evolutionary search in DisWOT automatically imitates good student from weak ones. In distillation phase, this searched student is distilled via teacher model and achieves superior gains.

training costs, our training-free DisWOT enjoys obvious differences than these methods and advantages in efficiency.

Compared to Other Training-free NAS. Table 3 clearly summarizes the differences between DisWOT and other zero-cost methods [1, 28, 36, 42, 58]. Moreover, DisWOT outperforms these methods on distillation performance prediction and boosting in our sufficient experiments dealing with diverse datasets and search spaces.

3. Methodology

Figure 3 provides an overview of the DisWOT framework, which is comprised of two main stages: optimal student network search and distillation with high-order knowledge. In the search stage, we employ the neural architecture search technique to obtain an optimal student network for a pre-defined teacher network. Notably, we propose a training-free proxy called DisWOT to accurately rank enormous student networks and prevent expensive evaluation processes with high efficiency. In the distillation stage, the searched student network is retrained with distillation to imitate high-order knowledge in the teacher network. We give the details of these two designs in the following sections.

3.1. Search for Optimal Student Network

We first present the training-free metrics we designed to score a student architecture, which indicates its final accuracy when distilled with a pre-defined teacher network. Then we depict the details of the evolutionary process to obtain an optimal student candidate.

Semantic Similarity Metric. The semantic information is meaningful for neural networks to perceive as humans. In distillation, the teacher network always has more convolutional operations than the student, resulting in a teacher feature map with a larger receptive field and greater richness of semantic information. In contrast to distiller designs to alleviate semantic gaps, we aim for train-free student architecture to better match the teacher model with computational constraints. We notice that the network with random initial weights also has some semantic localization capability. Thus, we start to analyze the localization performance of the randomly initialized teacher-student model. Specifically, we utilize Grad-CAM maps [78] to localize semantic object regions, which explains the model decisions using gradient information. Given a mini-batch of input images, we define the high-level feature map before the Global Average Pooling (GAP) layer of the teacher network T as $A_T \in \mathbf{R}^{B\times C_T\times H_T\times W_T}$, where B represents the batch size, C_T denotes the number of output channels, and H_T and W_T are the spatial dimensions. Additionally, we introduce $A_T^c \in \mathbf{R}^{N\times H_T\times W_T}$ as the c-th spatial map along the channel dimension. For the student network S_i, we have feature map $A_{S_i}^c \in \mathbf{R}^{B\times C_S\times H_S\times W_S}$ and spatial map $A_{S_i}^c \in \mathbf{R}^{B\times H_S\times W_S}$, respectively. To compute the Grad-CAM maps of the n-th class for both the teacher and student networks, we can use the following formulations:

$$G_T = \sum_{c=1}^{C_T} w_{n,c}^T A_T^c, \quad G_{S_i} = \sum_{c=1}^{C_S} w_{n,c}^S A_{S_i}^c, \tag{2}$$

where $w^T \in \mathbf{R}^{N\times C_T}$ and $w^S \in \mathbf{R}^{N\times C_S}$ are weights of the last fully-connected layer in the teacher and student network. N represents the number of classes. $w_{n,c}^T$ and $w_{n,c}^S$ refer to the element located in the n-th row and c-th column of weight matrices w^T and w^S, respectively. To quantify

the intersection of class-discriminative localization maps, we formulate semantic similarity metric $\mathcal{M}_s$ as the inter-correlation on the accumulated Grad-CAM maps for both teacher and student networks as follows:

$$\mathcal{G}^T = \frac{(G_T)\cdot(G_T)^\top}{\|(G_T)\cdot(G_T)^\top\|_2}, \mathcal{G}^S = \frac{(G_S)\cdot(G_S)^\top}{\|(G_S)\cdot(G_S)^\top\|_2}, \quad (3)$$

$$\mathcal{M}_s = \left\|\mathcal{G}^T - \mathcal{G}^{S_i}\right\|_2. \quad (4)$$

Relation Similarity Metric. The relationships between input samples are non-trivial for knowledge transfer. To reduce the teacher-student gap and improve the relation-distillation performance, we use the correlation matrix as the sample-wise metric to search for an optimal student network. For the random teacher network T and student network S_i with activation maps $A_T \in \mathbf{R}^{N\times C_T\times H_T\times W_T}$ and $A_S^i \in \mathbf{R}^{N\times C_i\times H_i\times W_i}$, the correlation matrix of the mini-batch samples in the teacher network is formulated as follows:

$$\mathcal{A}^T = \frac{(\tilde{A}_T)\cdot(\tilde{A}_T)^\top}{\left\|(\tilde{A}_T)\cdot(\tilde{A}_T^\top)\right\|_2}, \mathcal{A}^{S_i} = \frac{(\tilde{A}_S)\cdot(\tilde{A}_S)^\top}{\left\|(\tilde{A}_S)\cdot(\tilde{A}_S^\top)\right\|_2}, \quad (5)$$

where $\tilde{A}_T \in \mathbf{R}^{N\times CHW}$ is a reshaping of A_T, and M_T is a $N\times N$ matrix. Thus, the (i,j) entry in matrix C_T represents the similarity between the i-th and j-th images within the mini-batch. Based on this, the sample similarity metric $\mathcal{M}_r$ for a potential student model S_i is defined as follows:

$$\mathcal{M}_r = \left\|\mathcal{A}^T - \mathcal{A}^{S_i}\right\|_2. \quad (6)$$

Training-Free Evolutionary Search. Based on the above metric, we conduct a training-free evolutionary search algorithm to efficiently discover the optimal student α^* from search space $\mathcal{A}$, as:

$$\alpha^* = \arg\min_{\alpha\in\mathcal{A}}(\mathcal{M}_s + \mathcal{M}_r). \quad (7)$$

Theoretical Understanding. According to the VC theory [63], the classification error of the vanilla teacher-student network can be decomposed as follows:

$$R(f_s) - R(f_r) \le O\left(\frac{|\mathcal{F}_s|_C}{n^{\alpha_{sr}}}\right) + \epsilon_{sr}; R(f_t) - R(f_r) \le O\left(\frac{|\mathcal{F}_t|_C}{n^{\alpha_{tr}}}\right) + \epsilon_{tr}, \quad (8)$$

where $f_s \in \mathcal{F}_s$ is the student function, $f_t \in \mathcal{F}_t$ is the teacher function, and $f_r \in \mathcal{F}_r$ is the target function. R is the error. $O(\cdot)$ and ϵ_{sr} terms are the estimation and approximation error, respectively. $O(\cdot)$ is related to the statistical procedure when given the number of data points. In contrast, ϵ_{sr} is the approximation error of the student function class $\mathcal{F}_s$ for $f_r \in \mathcal{F}_r$. $|\cdot|_C$ is an function class capacity measure, and n is the number of data point. During distillation, the student network is supervised purely with the teacher network as follows:

$$R(f_s) - R(f_t) \le O\left(\frac{|\mathcal{F}_s|_C}{n^{\alpha_{st}}}\right) + \epsilon_{st}, \quad (9)$$

where α_{st} and ϵ_{st} are associated to student learning from teacher. By combining Equations 3.1 and 9, we obtain:

$$R(f_s) - R(f_r) \le O\left(\frac{|\mathcal{F}_t|_C}{n^{\alpha_{tr}}}\right) + \epsilon_{tr} + O\left(\frac{|\mathcal{F}_s|_C}{n^{\alpha_{st}}}\right) + \epsilon_{st}. \quad (10)$$

When student obtains gains in KDs, its upper bound of error in distillation is smaller than vanilla training, which satisfies the following inequality:

$$O\left(\frac{|\mathcal{F}_t|_C}{n^{\alpha_{tr}}} + \frac{|\mathcal{F}_s|_C}{n^{\alpha_{st}}}\right) + \epsilon_{tr} + \epsilon_{st} \le O\left(\frac{|\mathcal{F}_s|_C}{n^{\alpha_{sr}}}\right) + \epsilon_{sr}. \quad (11)$$

Based on the assumption in [22] that $\epsilon_{tr} + \epsilon_{st} \le \epsilon_{sr}$ holds consistently, we focus on minimizing $O\left(\frac{|\mathcal{F}_t|_C}{n^{\alpha_{tr}}} + \frac{|\mathcal{F}_s|_C}{n^{\alpha_{st}}}\right)$ to improve the distillation performance. As noted in Lopez-Paz et al [41], a better representation allows for a faster learning rate with a fixed amount of data. Hence, when there is a larger gap between the capacities of the student and teacher networks, the value of α_{st} tends to be lower. Thus we aim to search for an optimal student network that meets the requirement of $\alpha_{s_i t} \le \alpha_{s_o t}$, where s_i is all candidate student networks, and s_o is our searched student network. In this case, the inequality becomes more effective, and we improve the knowledge distillation by injecting a larger $\alpha_{s_o t}$. Specifically, we present the overall procedure for discovering optimal student in algorithm 1.

Effects of Search Strategies. We compare the evolution search algorithm and the random search algorithm in search space S_2 with the same number of iterations, as shown in Figure 4. We find that the evolution search algorithm can consistently find architectures with lower DisWOT, especially when the search space is relatively large, and the evolutionary search can explore better architectures.

3.2. Distillation with High-order Knowledge

In the distillation stage, teacher model T is employed to distill the optimal student network f_S. To verify the superiority of our search architecture, we adopt the existing distillers (e.g., KD) as the default distillation setting. In addition, we observe that the metrics we searched for actually serve as minimization optimization goals in the distillation process to transfer the teacher's privileged semantic and sample relational knowledge as the semantic distillation and sample distillation:

$$\mathcal{L}_{\mathcal{M}_s} = \frac{1}{c^2}\left\|\mathcal{G}^T - \mathcal{G}^S\right\|_2, \mathcal{L}_{\mathcal{M}_r} = \frac{1}{b^2}\left\|\mathcal{A}^T - \mathcal{A}^S\right\|_2, \quad (12)$$

Finally, we involve these advanced distillers in our framework, called DisWOT†. The total loss for DisWOT and

Algorithm 1 Evolution Search for DisWOT

Input: Search space $\mathcal{S}$, population $\mathcal{P}$, architecture constraints $\mathcal{C}$, max iteration $\mathcal{N}$, sample ratio r, sampled pool $\mathcal{Q}$, topk k, teacher network $\mathcal{T}$.
Output: Highest DisWOT score architecture.

$\mathcal{P}_0$:= Initialize population$(P_i, \mathcal{C})$;
sample pool $\mathcal{Q} := \emptyset$;
for $i = 1 : \mathcal{N}$ **do**
  Clear sample pool $\mathcal{Q} := \emptyset$;
  Randomly select $r \times \mathcal{P}$ subnets $\hat{P}_i \in \mathcal{P}$ to get $\mathcal{Q}$;
  Candidates $\{A_i\}_k$:= GetTopk($\mathcal{Q}$, k);
  Parent A_i := RandomSelect($\{A_i\}_k$);
  Mutate $\hat{P}_i$:= MUTATE(A_i);
  if $\hat{P}_i$ do not meet the constraints $\mathcal{C}$ **then**
    Do nothing;
  else
    Get DisWOT-Score z := DisWOT($\hat{P}_i$, $\mathcal{T}$);
    Append $\hat{P}_i$ to $\mathcal{P}$;
  end if
  Remove network of smallest DisWOT-score;
end for

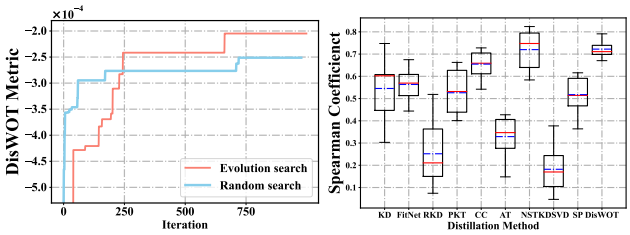

Figure 4. Left: comparison of random search and evolution search. Right: ranking correlation of different distillation methods on NAS-Bench-201.

Table 4. Spearman correlation ρ (%) on NAS-Bench-201.

Type	Method	ρ	Method	ρ
Zero-cost Proxies	Grad_Norm [1]	58.70±0.11	Synflow [58]	**74.61±0.08**
	SNIP [28]	58.17±0.15	Jacob [57]	73.42±0.03
	Fisher [1]	35.91±0.09	Zen-NAS [36]	41.36±0.06
	NWOT [42]	64.41±0.08	FLOPs [1]	63.38±0.06
KD-based Proxies	KD [22]	54.43±0.09	PKT [47]	52.65±0.09
	FitNets [55]	56.18±0.09	CC [48]	65.90±0.08
	SP [62]	51.24±0.08	NST [25]	72.35±0.09
	RKD [44]	25.71±0.17	DisWOT	**72.36±0.02**

DisWOT† as:

$$\mathcal{L}_{\text{DisWOT}} = \mathcal{L}_{CE}(f_S, Y) + \mathcal{L}_{KL}(f_S, f_T), \qquad \mathcal{L}_{\text{DisWOT}\dagger} = \mathcal{L}_{\text{DisWOT}} + \mathcal{L}_{\mathcal{M}_s} + \mathcal{L}_{\mathcal{M}_r}. \tag{13}$$

3.3. Bridging Distiller and Zero-proxy

In our DisWOT framework, we use the semantic and relational similarity metrics as a distillation performance predictor and distiller. In addition, DisWOT also enjoys good performance for vanilla performance predictions. Encouraged by this intriguing observation, we employ the knowledge function in of different KDs as zero-proxies and evaluate their ranking consistency with vanilla accuracy. As shown in Table 4, these KD-based zero-proxies enjoy competitive rankings with other NAS methods. Detailed results in Figure 4 illustrate that our DisWOT and NST [24] are the winners in the family of KD-based proxies. These attempts reveal the close connections between KD and NAS, and augment 10+ new universal proxies from the teacher-student learning perspective for training-free NAS research.

4. Experimental results

In this section, we present the experimental results of our DisWOT on different datasets. First, we describe the four datasets used in our experiments and three search spaces S_0, S_1, S_2 in Sec. 4.1. Then, we conduct a comprehensive set of experiments to evaluate the effectiveness of DisWOT.

4.1. Experimental Setup

We perform experiments on four datasets, namely CIFAR-10, CIFAR-100, ImageNet-16-120, and ImageNet-1k. In the search process, we only use one batch of training data to get the statistic at nearly no cost. Following previous works, our experiments are conducted on the following three search spaces:

Search Space S_0: Following cifar-ResNet [20], the search space consists of three residual blocks and is based on CIFAR-100 datasets. The depth of each residual block is searched in set {1,3,5,7}.

Search Space S_1: Following NAS-Bench-201 [14], this Darts-like search space is a cell-based search space consisting of stacked directed acyclic graphs. S_1 is conducted on CIFAR-10, CIFAR-100, and ImageNet-16-120 datasets.

Search Space S_2: Following NDS [53], this search space consists of residual and bottleneck blocks defined in ResNet. S_2 is based on CIFAR-100 and ImageNet-1k dataset.

4.2. Experiments on CIFAR-100

Implementation Details. We compare distillation gains with other zero-nas on search space S_1. In the search phase, we configure 48k evolution iters with 512 population sizes. In distillation, All searched student networks are trained via CRD's settings [59] with ResNet56 as the teacher model.

Distillation Results of Zero-cost Proxies. We conduct detailed experiments on other zero-cost proxies with different knowledge distillation methods. Note that we search the student network under constraints of 1M parameters. The results in Table 5 demonstrated that our proposed DisWOT achieved superior results compared with other zero-cost proxies with different knowledge distillation methods. The DisWOT outperforms its counterparts vanilla networks by around 2%, while achieving consistent improvements

Table 5. Distillation results (%) of different zero-cost proxies with knowledge distillation methods under 1M parameters.

Method	Random	FLOPs	Synflow	NWOT	DisWOT
Baseline	69.52	71.37	72.88	71.80	73.12
KD	70.45	72.13	73.72	72.57	74.73
FitNets	70.12	72.40	73.55	72.72	74.85
AT	70.16	72.97	73.52	72.08	74.50
SP	70.46	72.14	73.50	72.16	74.95
RKD	71.19	72.22	73.69	72.63	74.62
CRD	71.59	72.78	73.99	73.12	75.25

among different distillation methods, such as KD [22], FitNets [55], AT [74], SP [62], RKD [44], and CRD [60].

Table 6. Distillation results(%) of zero-cost proxies under {0.5,1,2}M parameters.

Param.	FLOPs	NWOT	DisWOT	DisWOT†
0.5M	69.88	70.38	72.89	**73.75**
1M	72.13	72.57	74.23	**75.25**
2M	73.27	73.86	75.95	**76.67**

Table 7. Distillation results(%) of zero-cost proxies under {50,100}M FLOPs on space S_1.

FLOPs	NWOT	Synflow	DisWOT	FLOPs	NWOT	Synflow	DisWOT
50M	63.19	64.28	65.98	100M	70.38	72.12	72.89

Table 8. Ranking correlation (%) of zero-cost proxies on S_0 space on CIFAR-100.

Method	Kendall's Tau	Spearman	Pearson
FLOPs [1]	51.61	72.92	76.40
Fisher [1]	62.86	81.37	20.90
Grad_Norm [1]	63.75	82.35	39.35
SNIP [28]	67.22	85.07	51.09
NWOT [42]	31.87	45.66	48.99
DisWOT (ours)	**73.98**	**91.38**	**84.83**

Analysis on Varying Parameter Constraints. We analyze the performance of student models under different parameter constraints obtained by DisWOT on CIFAR-100. As shown in the Table 7, we compared our method with two zero proxies, a.k.a. FLOPs [1] and NWOT [42], under the parameter constraints of 0.5, 1, and 2M, respectively, and the results demonstrate that our method still achieves excellent results. As shown in Tab. 7, DisWOT also outperforms previous SOTA methods with 0.8%∼1.7%↑ gains under same FLOPs constraints,

Ranking Correlation with Distill Accuracy. Based on search space S_0, we perform vanilla training and distillation for each candidate with CRD's settings [59]. Then, we collect these vanilla results as GT and analyze the different zero-proxy's correlation with them. As shown in Table 8, the results illustrate that our DisWOT achieves higher than Fisher, GradNorm, SNIP,FLOPs, and NWOT by a large margin, and achieve results that are on par with the best zero-cost proxy, a.k.a. Zen-NAS and Synflow, on Kendall's Tau, Pearson, and Spearman coefficient.

4.3. Experiments on NAS-Bench-201

Implementation Details. For search trials, we first adopt ResNet110/56 as the teachers and then Conduct an evolution search with the DisWOT metric and get the best student network. We randomly sampled 50 candidate architectures to evaluate sequencing consistency. The distillation settings are the same as the Sec.4.2

Comparison results As shown in Table 9, some training-free methods can achieve good results with much faster speedups, such as NWOT and TE-NAS. Our proposed method DisWOT achieves a speedup ratio of 180×, where if semantic similarity metric is removed, we can achieve a 300× speedup ratio at the expense of some accuracy.

4.4. Experiments on ImageNet

Implementation Details. We searched the ResNet18 level network regarding the search space in NDS [53]. Specifically, we limit the number of parameters to less than 13M and the depth of the network to up to 20 layers and find the optimal network by evolution algorithm with the DisWOT metric. As shown in Table 10, guided by three different sizes of networks, we used DisWOT to find the optimal student network. We trained the student network obtained by the search using the distillation strategy in DisWOT. Implementation details are available in supplementary materials.

Comparison Results. Table 10 reports the performance of DisWOT on ImageNet with ResNet34/50 as teacher network. The results demonstrate that the student architecture of the ResNet18-level obtained by DisWOT under different teacher guidance and using different distillation strategies yielded significantly better results than its counterparts.

4.5. Ablation Studies of DisWOT

We perform ablation experiments to verify the validity of each component of DisWOT in search space S_0. As shown in Table 11, for semantic knowledge, similarity matrix obtains a more robust ranking improvement than simple FitNet [54]. For $\mathcal{M}_r$, similarity matrix performs better on relational knowledge than RKD [45]. DisWOT integrates semantic and relational knowledge to obtain an additional ranking improvement than stand-alone scores. The weight initialization scheme plays an important role in zero-proxy.

Table 9. Distillation results on CIFAR-10, CIFAR-100, and ImageNet-16 in NAS-Bench-201 [13]. Dis. Acc. (%) represents the accuracy of the searched architecture after distillation training. Time (s) denotes the time cost (GPU-seconds) during the search phase. The results of NWOT and TE-NAS come from their original papers. Our DisWOT achieves competitive results with the lowest costs.

Type	Model	CIFAR-10			CIFAR-100			ImageNet-16-120		
		Dis. Acc(%)	Time (s)	Speed-up	Dis.Acc(%)	Time (s)	Speed-up	Dis. Acc(%)	Time (s)	Speed-up
Multi-trial	RS	93.63	216K	1.0×	71.28	460K	1.0×	44.88	1M	1.0×
	RL [3]	92.83	216K	1.0×	71.71	460K	1.0×	44.35	1M	1.0×
	BOHB [17]	93.49	216K	1.0×	70.84	460K	1.0×	44.33	1M	1.0×
	RSPS [34]	91.67	10K	21.6×	57.99	46K	21.6×	36.87	104K	9.6×
Weight-sharing	GDAS [15]	93.39	22K	12.0×	70.70	39K	11.7×	42.35	130K	7.7×
	DARTS [38]	89.22	23K	9.4×	66.24	80K	5.8×	43.18	110K	9.1×
Training-free	NWOT [42]	93.73	2.2K	100×	73.31	4.6K	100×	45.43	10K	100×
	TE-NAS [9]	93.92	2.2K	100×	71.24	4.6K	100×	44.38	10K	100×
DisWOT	$\mathcal{M}_s$ & $\mathcal{M}_r$	93.55	1.2K	180×	74.21	9.2K	180×	47.30	20K	180×
	$\mathcal{M}_r$	93.49	0.72K	**300**×	73.62	18.4K	**300**×	45.63	40K	**300**×

Table 10. The accuracy (%) of ResNet18 on ImageNet-1k with various teachers. Results of other KD methods refer to the papers of CRD [60] and ESKD [10]. ATKD $_{A_{R34}}$ [43] denotes ResNet34 used as the assistant teacher. N/A means no available results. Our DisWOT obtains better performance than other methods and improves students' performance positively correlated with that of the teacher.

Teacher	Student	Acc.	Teacher	Student	KD [22]	ESKD [10]	ATKD$_{A_{R18}}$ [43]	ONE [27]	DML [75]	CRD [60]	DisWOT
ResNet34	ResNet18	Top-1	73.40	69.75	70.66	70.89	70.78	70.55	71.03	71.17	**72.08**
		Top-5	91.42	89.07	89.88	90.06	89.99	89.59	90.28	90.32	**90.38**
Teacher	Student	Acc.	Teacher	Student	KD [22]	ATKD$_{A_{R18}}$ [43]	ATKD$_{A_{R34}}$ [43]	Seq. ESKD [10]	ESKD [10]	SRRL [72]	DisWOT
ResNet50	ResNet18	Top-1	76.16	69.75	70.68	70.65	70.85	70.65	70.95	71.20	**72.30**
		Top-5	92.86	89.07	N/A	N/A	N/A	N/A	N/A	N/A	**90.51**

Table 11. Spearman correlation ("mean±std") of DisWOT on search space S_0.

Knowledge	Metric	Spearman (%)
$\mathcal{M}_s$	FitNets [54]	64.06±6.11
$\mathcal{M}_s$	Similarity matrix	73.68±5.45
$\mathcal{M}_r$	RKD [59]	13.52±11.51
$\mathcal{M}_r$	Similarity matrix	72.36±3.42
$\mathcal{M}_s$ & $\mathcal{M}_r$	Similarity matrix	**77.51±2.76**

Table 12. "mean±std %" Spearman of proxies via Kaiming and Gaussian initialization on search space S_0 and NAS-Bench-201 with various seeds.

Space	Initial	Fisher	GradNorm	NWOT	DisWOT
S_0	Kaim.	81.37±0.01	82.35±0.01	45.66±0.05	84.08±0.03
	Gauss.	80.99±0.01	75.50±0.01	45.36±0.03	91.38±0.03
NB-201	Kaim.	54.63±0.15	58.70±0.11	64.41±0.08	65.57±0.02
	Gauss.	45.91±0.09	45.70±0.11	62.24±0.07	72.36±0.02

We verify the effect of the initialization strategy of the network on the ranking consistency. The results in Tab. 12 demonstrate that the Gaussian initialization strategy is detrimental to $\mathcal{M}_s$, but beneficial to $\mathcal{M}_r$.

5. Conclusion

In this paper, we present DisWOT, a new teacher-aware student architecture search without training framework for distillation. Based on key observations about the difference between vanilla and distillation accuracy, DisWOT measures the new zero-cost proxy conditioned on the similarity of feature semantics and sample relations between random-initialized teacher-student network. Then, DisWOT search for the best student architectures for the given teacher using an evolutionary algorithm with these metrics. Thorough evaluations are performed on diverse datasets and search spaces, and DisWOT achieves significant performance gains in various neural networks with at least 180× training acceleration. We experimentally and theoretically explained the relationship between similarity difference and distillation performance. In addition, we also extend DisWOT to new distillers and general zero proxy to predict the performance of models. By doing this, we bridge the higher-order knowledge bewteen distillation and network architecture search. This approach represents an elegant and practical solution, which we hope will inspire future research on knowledge distillation and architecture search design.

Limitations. Following most zero-cost NAS, we evaluate DisWOT in classification tasks. In the future work, we will make efforts to expand the DisWOT for downstream tasks (e.g., object detection and semantic segmentation).

References

[1] Mohamed S Abdelfattah, Abhinav Mehrotra, Łukasz Dudziak, and Nicholas Donald Lane. Zero-cost proxies for lightweight nas. In ICLR, 2020. 3, 4, 6, 7

[2] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In CVPR, 2019. 1

[3] Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. In ICLR, 2017. 8

[4] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In CVPR, 2022. 1

[5] Tom B. Brown, Benjamin Mann, Nick Ryderand Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M, Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. arXiv preprint, arXiv:2005.14165, 2020. 1

[6] Yun-Hao Cao and Jianxin Wu. A random cnn sees objects: One inductive bias of cnn and its applications. In AAAI, 2022. 2

[7] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In CVPR, 2022. 1

[8] Kunlong Chen, Liu Yang, Yitian Chen, Kunjin Chen, Yidan Xu, and Lujun Li. Gp-nas-ensemble: a model for the nas performance prediction. In CVPRW, 2022. 3

[9] Wuyang Chen, Xinyu Gong, and Zhangyang Wang. Neural architecture search on imagenet in four gpu hours: A theoretically inspired perspective. In ICLR, 2020. 8

[10] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In ICCV, 2019. 1, 3, 8

[11] Peijie Dong, Xin Niu, Lujun Li, Zhiliang Tian, Xiaodong Wang, Zimian Wei, Hengyue Pan, and Dongsheng Li. Rd-nas: Enhancing one-shot supernet ranking ability via ranking distillation from zero-cost proxies. arXiv preprint arXiv:2301.09850, 2023. 3

[12] Peijie Dong, Xin Niu, Lujun Li, Linzhen Xie, Wenbin Zou, Tian Ye, Zimian Wei, and Hengyue Pan. Prior-guided one-shot neural architecture search. arXiv preprint arXiv:2206.13329, 2022. 3

[13] Xuanyi Dong and Yi Yang. Nas-bench-201: Extending the scope of reproducible neural architecture search. In ICLR, 2019. 8

[14] Xuanyi Dong and Yi Yang. Searching for a robust neural architecture in four gpu hours. In CVPR, 2019. 2, 6

[15] Xuanyi Dong and Yezhou Yang. Searching for a robust neural architecture in four gpu hours. CVPR, 2019. 8

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 1

[17] Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. In ICML, 2018. 8

[18] Jindong Gu and Volker Tresp. Search for better students to learn distilled knowledge. In ECAI, 2020. 3

[19] 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. arXiv preprint arXiv:1904.00420, 2019. 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016. 6

[21] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In AAAI, 2019. 1

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015. 1, 5, 6, 7, 8

[23] Yiming Hu, Xingang Wang, Lujun Li, and Qingyi Gu. Improving one-shot nas with shrinking-and-expanding supernet. Pattern Recognition, 2021. 3

[24] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219, 2017. 1, 3, 6

[25] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. arXiv:1707.01219, 2017. 6

[26] Jangho Kim, SeoungUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In NeurIPS, 2018. 1

[27] Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In NeurIPS, 2018. 8

[28] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip Torr. Snip: Single-shot network pruning based on connection sensitivity. In ICLR, 2018. 3, 4, 6, 7

[29] Changlin Li, Jiefeng Peng, Liuchun Yuan, Guangrun Wang, Xiaodan Liang, Liang Lin, and Xiaojun Chang. Block-wisely supervised neural architecture search with knowledge distillation. CVPR, 2020. 3

[30] Lujun Li. Self-regulated feature learning via teacher-free feature distillation. In ECCV, 2022. 3

[31] Lujun Li and Zhe Jin. Shadow knowledge distillation: Bridging offline and online knowledge transfer. In NeuIPS, 2022. 3

[32] Lujun Li, Liang Shiuan-Ni, Ya Yang, and Zhe Jin. Boosting online feature transfer via separable feature fusion. In IJCNN, 2022. 3

[33] Lujun Li, Liang Shiuan-Ni, Ya Yang, and Zhe Jin. Teacher-free distillation via regularizing intermediate representation. In IJCNN, 2022. 3

[34] Liam Li and Ameet S. Talwalkar. Random search and reproducibility for neural architecture search. ArXiv, 2019. 8

[35] Lujun Li, Yikai Wang, Anbang Yao, Yi Qian, Xiao Zhou, and Ke He. Explicit connection distillation. In ICLR, 2020. 3

[36] Ming Lin, Pichao Wang, Zhenhong Sun, Hesen Chen, Xiuyu Sun, Qi Qian, Hao Li, and Rong Jin. Zen-nas: A zero-shot nas for high-performance image recognition. 2021. 3, 4, 6

[37] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. Knowledge distillation via the target-aware transformer. In CVPR, 2022. 2

[38] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architecture search. In ICLR. 8

[39] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. arXiv preprint arXiv:1806.09055, 2018. 3

[40] Yu Liu, Xuhui Jia, Mingxing Tan, Raviteja Vemulapalli, Yukun Zhu, Bradley Green, and Xiaogang Wang. Search to distill: Pearls are everywhere but not the eyes. In CVPR, 2020. 1, 3

[41] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. arXiv preprint arXiv:1511.03643, 2015. 5

[42] Joe Mellor, Jack Turner, Amos Storkey, and Elliot J Crowley. Neural architecture search without training. In ICML, 2021. 3, 4, 6, 7, 8

[43] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In AAAI, 2020. 1, 3, 8

[44] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In CVPR, 2019. 6, 7

[45] Wonpyo Park, Yan Lu, Minsu Cho, and Dongju Kim. Relational knowledge distillation. In CVPR, 2019. 3, 7

[46] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In ECCV, 2018. 3

[47] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In ECCV, 2018. 6

[48] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In ICCV, 2019. 6

[49] Baoyun Peng, Xiao Jin, Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Yu Liu, Dong-sheng Li, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In ICCV, 2019. 3

[50] Houwen Peng, Hao Du, Hongyuan Yu, Qi Li, Jing Liao, and Jianlong Fu. Cream of the crop: Distilling prioritized paths for one-shot neural architecture search. NeurIPS, 2020. 3

[51] Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. In ICML, 2018. 3

[52] Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In AAAI, 2022. 1

[53] Ilija Radosavovic, Raj Prateek Kosaraju, Ross B. Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. CVPR, 2020. 6, 7

[54] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In ICLR, 2015. 1, 3, 7, 8

[55] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. ICLR, 2015. 6, 7

[56] R. R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, D. Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. 2019. 2

[57] Suraj Srinivas and François Fleuret. Knowledge transfer with jacobian matching. In ICML, 2018. 6

[58] Hidenori Tanaka, Daniel Kunin, Daniel L Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. NeurIPS, 2020. 3, 4, 6

[59] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In ICLR, 2020. 3, 6, 7, 8

[60] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In ICLR, 2020. 7, 8

[61] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In ICCV, 2019. 1

[62] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In ICCV, 2019. 6, 7

[63] Vladimir Vapnik. Statistical learning theory. 1998. 5

[64] Likang Wang and Lei Chen. Dionysus: Recovering scene structures by dividing into semantic pieces. 1

[65] Likang Wang and Lei Chen. Ftso: Effective nas via first topology second operator. 2023. 1

[66] Likang Wang, Yue Gong, Xinjun Ma, Qirui Wang, Kaixuan Zhou, and Lei Chen. Is-mvsnet: Importance sampling-based mvsnet. In ECCV, 2022. 1

[67] Likang Wang, Yue Gong, Qirui Wang, Kaixuan Zhou, and Lei Chen. Flora: dual-frequency loss-compensated real-time monocular 3d video reconstruction. In AAAI, 2023. 1

[68] Zimian Wei, Hengyue Pan, Lujun Li Li, Menglong Lu, Xin Niu, Peijie Dong, and Dongsheng Li. Convformer: Closing the gap between cnn and vision transformers. arXiv preprint arXiv:2209.07738, 2022. 1

[69] 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

[70] Liu Xiaolong, Li Lujun, Li Chao, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching. In ICLR, 2023. 3

[71] Jingjing Xu, Liang Zhao, Junyang Lin, Rundong Gao, Xu Sun, and Hongxia Yang. Knas: Green neural architecture search. In ICML, 2021. 3

[72] Jing Yang, Brais Martinez, Adrian Bulat, Georgios Tzimiropoulos, et al. Knowledge distillation via softmax regression representation learning. ICLR, 2021. 8

[73] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In ICLR, 2017. 1, 3

[74] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. ICLR, 2017. 7

[75] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In CVPR, 2018. 8

[76] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In CVPR, 2022. 1

[77] Xiawu Zheng, Xiang Fei, Lei Zhang, Chenglin Wu, Fei Chao, Jianzhuang Liu, Wei Zeng, Yonghong Tian, and Rongrong Ji. Neural architecture search with representation mutual information. CVPR, 2022. 3

[78] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In CVPR, 2016. 4

[79] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In CVPR, 2018. 3