UniADS: Universal Architecture-Distiller Search for Distillation Gap

Liming Lu\textsuperscript{1*}, Zhenghan Chen\textsuperscript{2*}, Xiaoyu Lu\textsuperscript{1†}, Yihang Rao\textsuperscript{1}, Lujun Li\textsuperscript{3}, Shuchao Pang\textsuperscript{1,4†}

\textsuperscript{1}School of Cyber Science and Engineering, Nanjing University of Science and Technology
\textsuperscript{2}Peking University
\textsuperscript{3}HKUST
\textsuperscript{4}School of Computing, Macquarie University

\{luliming, xiaoyu.lu, pangshuchao\}@njust.edu.cn, 1979282882@pku.edu.cn, lilujunai@gmail.com

Abstract

In this paper, we present UniADS, the first Universal Architecture-Distiller Search framework for co-optimizing student architecture and distillation policies. Teacher-student distillation gap limits the distillation gains. Previous approaches seek to discover the ideal student architecture while ignoring distillation settings. In UniADS, we construct a comprehensive search space encompassing an architectural search for student models, knowledge transformations in distillation strategies, distance functions, loss weights, and other vital settings. To efficiently explore the search space, we utilize the NSGA-II genetic algorithm for better crossover and mutation configurations and employ the Successive Halving algorithm for search space pruning, resulting in improved search efficiency and promising results. Extensive experiments are performed on different teacher-student pairs using CIFAR-100 and ImageNet datasets. The experimental results consistently demonstrate the superiority of our method over existing approaches. Furthermore, we provide a detailed analysis of the search results, examining the impact of each variable and extracting valuable insights and practical guidance for distillation design and implementation.

Introduction

Knowledge distillation (KD) methods (Hinton, Vinyals, and Dean 2015; Romero et al. 2015) have emerged as powerful techniques for model compression and transfer learning in the field of deep learning. These methods aim to transfer knowledge from a large, well-performing model, known as the teacher model, to a smaller, more compact model, referred to as the student model. By distilling the knowledge from the teacher model, these methods offer a pathway to achieve high accuracy while reducing the computational complexity and memory requirements of deep neural networks, making them more suitable for resource-constrained environments. Despite the progress made in KD designs, the distillation gap (Huang et al. 2022; Mirzadeh et al. 2020) between teacher and student models limits the improvement. While KDs have shown promise in transferring knowledge from large teacher models to smaller student models, the effectiveness of this process is hindered by the inherent disparity between the two models. Larger and more accurate teacher models often exhibit over-confidence and struggle to improve the performance of the student models effectively (Zhou et al. 2021). This distillation gap poses a challenge in achieving optimal knowledge transfer and limits the overall performance gains.

To address this issue, existing methods have explored alternative approaches, such as assistant models (Mirzadeh et al. 2020) and distillation-aware architecture search (Liu et al. 2020). Assistant models attempt to bridge the gap by introducing an additional model into the distillation process. This model acts as an intermediary between the teacher and student models, helping to facilitate knowledge transfer. However, the use of assistant models often requires additional training budgets and can result in increased computational costs. Another approach to mitigate the distillation gap is architecture search, which focuses on discovering optimal student architectures specifically for knowledge distillation. These methods aim to improve the knowledge transfer process and reduce the disparity between teacher and student models by searching for architectures that are well-suited for distillation. These approaches often use strategies like reinforcement learning, evolutionary algorithms, or Bayesian optimization to search for architectures that maximize the performance of the student model under the distillation frame-
work. However, existing student architecture searches often solely focus on the architectural dimension of the search and overlook the impact of distiller settings. Distiller settings refer to the specific choices made in the process of distillation, such as knowledge transformations, distance functions, loss weights, and other important settings that govern the knowledge transfer process. Distiller settings play a crucial role in shaping the interaction between the teacher and student models during the distillation process. The choice of knowledge transformations, such as attention transfer or feature micking, can significantly impact the transfer of knowledge from the teacher to the student. Similarly, the selection of an appropriate distance function and loss weights can influence the alignment between the teacher and student predictions, ensuring that the student model captures the essential information from the teacher model. By disregarding the effects of distiller settings, architecture search methods may produce architectures that are not fully compatible with the distillation strategies. This mismatch can result in inefficient knowledge transfer, inadequate teacher knowledge utilization, and, ultimately, sub-optimal distillation performance.

To address the challenges mentioned, we introduce UniADS, an innovative automated search framework specifically designed to tackle the joint optimization of architectural dimensions and distiller settings in the knowledge distillation process. UniADS offers a comprehensive and efficient approach by exploring the design space, encompassing various architectural choices, knowledge transformations, distance functions, loss weights, and other crucial settings. In UniADS, the search algorithm is vital in discovering optimal combinations of architectural dimensions and distiller settings. We employ the NSGA-II multi-objective optimization algorithm with accuracy-performance trade-offs as multiple conflicting objectives. By considering the evaluation results of different configurations, NSGA-II performs crossover and mutation operations to generate new candidate solutions. This iterative process aims to improve the overall performance of the student model while maintaining a diverse set of high-quality solutions. Furthermore, UniADS incorporates an acceleration strategy known as the Successive Halving algorithm. The Successive Halving algorithm is a progressive elimination strategy that efficiently prunes the search space by iteratively discarding underperforming configurations. This approach significantly reduces the computational burden and speeds up the distiller search process. In fact, our acceleration strategy achieves a remarkable 40-fold acceleration during the distiller search, enabling faster exploration of the design space and more efficient identification of promising architectures and distiller settings. By incorporating distiller settings into the search space, UniADS enables researchers to explore a broader range of possibilities. This holistic approach ensures that the discovered architectures are not only well-suited for distillation but also compatible with effective distiller settings. UniADS aims to achieve improved distillation results and a reduced gap between the teacher and student models by jointly optimizing architectural dimensions and distiller settings. This comprehensive optimization process maximizes the utilization of the teacher’s knowledge. It enhances the transfer of information to the student model, leading to compact and efficient student models that closely match the performance of their larger teacher counterparts.

Our approach, UniADS, offers several key advantages over traditional KD methods. It effectively reduces the teacher-student gap by introducing a general distiller search space and employing an adaptive evolutionary search technique. UniADS is efficient, eliminating the need for laborious manual tuning and increasing efficiency systematically. It also provides valuable insights by analyzing advanced distillation designs, offering practical guidance for implementation. In summary, UniADS presents a new direction for automated KD, enhancing effectiveness, improving efficiency, and facilitating future research in this field.

• To address the architecture and capability gaps between teachers and students, we introduce UniADS, a pioneering auto-search framework that evolves optimal distillers by leveraging current knowledge. This framework starts from fundamental concepts and rapidly incorporates advanced insights to develop new distillation approaches.
• We construct a comprehensive distiller search space that includes the architectural search for student models, knowledge transformations in distillation strategies, distance functions, loss weights, and other essential settings. Moreover, we utilize the NSGA-II genetic algorithm to identify and combine optimal configurations and employ the Successive Halving algorithm for efficient search space pruning.
• We conducted extensive experiments on the CIFAR-100 and ImageNet datasets with various teacher-student pairs. The experimental results consistently show that our method surpasses traditional KD techniques. Specifically, ResNet-18 enhanced with UniADS achieves a Top-1 accuracy of 72.51% on ImageNet, which is an improvement of 1.83% over conventional KD methods.

Related Work

Knowledge Distillation. Knowledge distillation (Hinton, Vinyals, and Dean 2015) transfers knowledge from a complex, large teacher model to a more efficient, smaller student model, aiming to equip the student with comparable performance. Attention-based techniques, such as AT (Zagoruyko and Komodakis 2017), enhance the student model’s performance by aligning attention maps, enabling a focus on pertinent input features. RKD (Park et al. 2019) further enriches this process by leveraging the teacher’s relational hints, bolstering the student’s training. The ‘distillation gap’—the performance disparity between teacher and student models—presents a persistent challenge. To narrow this gap, strategies like employing assistant teachers (Mirzadeh et al. 2020), architecture search, and tailored KD designs, including transformations (Huang et al. 2022), distance functions (Shu et al. 2021), and weight adjustments (Liu et al. 2022), have been proposed. For instance, ATKD (Mirzadeh et al. 2020) uses an intermediate student model as a bridge, facilitating knowledge transfer to the final student model by capturing a mid-level abstraction. In contrast, our UniADS method innovates by automating the search for optimal distillation strategies without necessitating additional architectural
Table 1: Specific operations in UniADS.

| Architecture | Depth, Width | Depth-{1,2,3} values:1,3,5,7; Width-ratio: 0.5,1,1.5,2 |
| Distiller | Transform | Distance | Weight |
| | | | | batchnorm, scale, multi — scale, local, batch, channel, drop, satt, natt, catt, mask, no no-norm loss: smooth $\ell_1$, $\ell_1$, $\ell_2$, $\ell_{KL}$, $\ell_{hard}$; norm loss: $\ell_{Cosine}$, $\ell_{Pearson}$, $\ell_{Correlation}$ weight values: 0.01,...,100; $\tau$ values:1,4,8,16 |

The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Methodology

In this section, we first present the design of our distillation search space, detailing the search methodology, acceleration techniques, and the objectives for fitting. Subsequently, we analyze the outcomes of the search and offer some practical guidelines. Lastly, we delve into the analysis of the student models distilled through UniADS, exploring their application across various distillation scenarios. The workflow of our approach is depicted in Figure 2.

Universal Architecture-Distiller Search Space

Search space structure. In KD, the student student $S$ is distilled with the fixed teacher $T$ by minimizing:

$$L_{KD} = \tau^2 \times W \times D(A_S) / \tau, \quad (1)$$

where $W$ is the loss weights factor, $\tau$ is the temperature factor, $T$ is transformations, $D(\cdot, \cdot)$ is distance function measuring the knowledge difference. $A$ are outputs (e.g., features, embeddings, and logits) of the teacher-student. Our search space consists of different types of operations (see Table 1) in transformations, distance functions, and loss weights of the distiller and depth & width-ratio of student architecture. Following this general KD formulation, we select operators normalized in different dimensions (e.g., batchnorm, norm $H,W,C,N$), various types of activation functions (e.g., exp, relu, tanh, sigmoid, pow2), multi-scale process and spatial-wise/channel-wise mask transforms and other advanced operations in the knowledge transformation. Our distance function options include smooth $\ell_1$, $\ell_1$, $\ell_2$, $\ell_{KL}$, $\ell_{hard}$, $\ell_{Cosine}$, $\ell_{Pearson}$ and $\ell_{Correlation}$ distance. Options in the loss weight part include various values for loss factors and temperature factors. Then, we use a computation graph to represent each candidate, in which the input nodes are different types of knowledge and the intermediate nodes are primitive operations.

Accuracy-Efficiency Trade-offs Search

In our architecture-distiller search aiming to optimize accuracy and model parameters simultaneously, NSGA-II is applied as a multi-objective genetic algorithm. It explores the architectural space, generating diverse architectures that trade-off between accuracy and model parameters. To accurately evaluate each distiller and reduce the distillation gap, we include test loss and model parameters as the multi-objectives. Specifically, we conduct a gradient-free evolutionary search algorithm to efficiently discover the optimal distiller $\alpha^*$ from search space $A$, as:

$$\alpha^* = \arg \min_{\alpha \in A}(L_{CE}(f(A_S), Y) + \|Param.(A_S) - C\|), \quad (2)$$
where $L_{CE}$ is the regular cross-entropy objective with labels $Y$. $C$ is a pre-defined constraint of the model parameters.

The algorithm initializes a population, evaluates architectures based on both objectives, performs non-dominated sorting and crowding distance assignment, selects architectures for the next generation, applies genetic operators, and replaces individuals. This process continues until a termination condition is met. By considering accuracy and model parameters as multi-target objectives, NSGA-II efficiently discovers architectures that strike a balance between the two objectives in the distillation process.

**Successive Halving Search Acceleration**

Successive Halving is employed as the underlying search acceleration technique. It starts with a large population of candidate distillation architectures. The process is then carried out in several stages, each consisting of the following steps:

1. Initialization: The search begins with a large population of candidate architectures. 2. Evaluation and Loss-Rejection Protocol: Each architecture in the population is trained and evaluated. However, the loss-rejection protocol is employed to accelerate the search and filter out unpromising distillers. Candidates with excessive loss values or collapsed optimization, indicating poor performance, are filtered out and eliminated from further consideration. 3. Search Space Shrinkage: As Successive Halving proceeds, the search space is systematically narrowed. Operations tied to frequent loss rejections and poor performers see reduced sampling probabilities, concentrating computational resources on more promising architectures. 4. Resource Reallocation: Resources are reallocated to the remaining candidates after the less promising distillers are discarded, allowing the search to advance with a focused subset. 5. Iteration: The cycle of evaluation, shrinkage, and reallocation (Steps 2-4) is repeated, each iteration refining the pool of architectures by removing the least successful ones. 6. Selection: The search culminates in the selection of the final architecture, based on its performance in the concluding iteration or upon meeting a predefined termination criterion.

In conjunction with the mentioned strategies, the Successive Halving technique allows for an efficient and accelerated search for optimal distillation architectures. The loss-rejection protocol filters out unpromising candidates, the search space shrinkage focuses computational efforts on more promising architectures, and Successive Halving progressively allocates resources to the remaining architectures. This combination of strategies helps streamline the search process, reducing the computational burden and enabling the identification of optimal distillation architectures more quickly.

**Search Results Analysis**

Table 2 presents searched distillers for different models. Based on these results, some practical guidance for KD designs can be summarized as:

- **Transform $T$:** Some normalized-based methods are “batchnorm”, “tanh”, and “norm”. These methods are applied to manipulate the input data before it is used for distillation. For example, “batchnorm” refers to batch normalization, which normalizes the input data by subtracting the mean and dividing by the standard deviation.
- **Distance metric $D(\cdot, \cdot)$:** The most frequently occurring metrics are “cos”, “kl” and “l2”. This suggests that some
Table 2: Top-1 accuracies (%) of different methods on CIFAR-100. KD denotes random student architecture distilled by the search by KD (Hinton, Vinyals, and Dean 2015). DAS denotes DisWOT search student architecture of DisWOT searched distilled by KD. UniADS and detailed architecture-distillation search profiles are shown in Table 2.

Experiments

In this section, we evaluate our method on CIFAR-100 and ImageNet datasets and compare its performance with existing knowledge distillation methods. For fair comparisons, we use the public codes and adopt the same training and data preprocessing settings for all reference methods. We also perform various ablative experiments to analyze the key designs of our UniADS. All experiments are conducted with PyTorch (Paszke et al. 2019) for 3 separate runs. Full implementation details are available in the supplementary materials.

Experiments on CIFAR-100

Dataset. CIFAR-100 (Krizhevsky and Hinton 2009), containing 50,000 training images and 10,000 test images with 100 classes, is the most popular classification dataset for evaluating the performance of knowledge distillation methods.

Implementation. We choose prevalent ResNets (He et al. 2016) of different depths and WRNs (Zagoruyko and Komodakis 2016) of different widths as the student networks to conduct experiments. All teacher-student networks are trained with the typical training setting of 200 epochs, following the original papers. During the distiller search phase, we adopt a basic tree structure search space and training acceleration settings, including 24 early-stop training epochs. Our UniADS search performs 200 iterations for each teacher-student pair. During the distillation phase, all teacher-student networks are trained using typical training settings, with a training epoch of 240. We set the batch size to 128 and use a standard SGD optimizer. The learning rate is initialized to 0.1 and decays by 0.1 at 100 and 150 epochs.

Compared to the distillation-aware architecture search method. Table 2 presents the average results obtained from Random Search, Architecture Search, as well as our proposed methods UniADS. Specifically, focusing on the student model with ResNet backbones, we observe substantial absolute accuracy gains ranging from 1.69% to 2.71% when utilizing UniADS. This observation highlights the effectiveness of our methods in improving the performance of ResNets while accommodating various constraints. Although different teacher models yield only modest accuracy gains when employing the same knowledge distillation strategy, our UniADS approach demonstrates significant improvement over both architectural and random searches. This highlights the superiority of joint architectural-distiller search. These findings underscore UniADS’s effectiveness in optimizing student model performance. It showcases UniADS’s ability to
Achieve substantial accuracy improvements, particularly with ResNet-based models, while also illustrating its superiority over traditional architectural and random search approaches.

**Comparison with existing KD methods with individual distiller search.** Table 3 presents a comparative analysis of our UniADS with other methods (e.g., FitNets (Romero et al. 2015), AT (Zagoruyko and Komodakis 2017), SP (Tung and Mori 2019), CC (Peng et al. 2019), RKD (Park et al. 2019), CRD (Tian, Krishnan, and Isola 2020), DIST (Huang et al. 2022), SRRL (Jing Yang 2021) and SemCKD (Liu, Liu, and Huang 2022)). We report Top-1 “mean (std)” accuracies (%) for UniADS over 3 runs under the same CRD (Tian, Krishnan, and Isola 2020) training settings.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tea_Acc</td>
<td>75.61</td>
<td>74.31</td>
<td>74.31</td>
<td>79.42</td>
<td>74.64</td>
<td>74.64</td>
<td>79.42</td>
<td>71.82</td>
<td>70.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stu_Acc</td>
<td>73.26</td>
<td>69.06</td>
<td>71.14</td>
<td>72.50</td>
<td>70.36</td>
<td>64.60</td>
<td>74.45</td>
<td>74.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KD</td>
<td>74.92</td>
<td>70.67</td>
<td>73.08</td>
<td>73.33</td>
<td>72.98</td>
<td>67.37</td>
<td>74.45</td>
<td>74.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FitNet</td>
<td>73.58</td>
<td>68.99</td>
<td>71.06</td>
<td>73.50</td>
<td>71.02</td>
<td>64.14</td>
<td>73.54</td>
<td>73.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td>74.08</td>
<td>70.22</td>
<td>72.31</td>
<td>73.44</td>
<td>71.43</td>
<td>59.40</td>
<td>72.73</td>
<td>73.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP</td>
<td>73.83</td>
<td>70.04</td>
<td>72.69</td>
<td>72.94</td>
<td>72.58</td>
<td>66.30</td>
<td>74.56</td>
<td>74.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>73.56</td>
<td>69.48</td>
<td>71.48</td>
<td>72.97</td>
<td>70.71</td>
<td>64.86</td>
<td>71.29</td>
<td>71.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RKD</td>
<td>73.35</td>
<td>69.25</td>
<td>71.82</td>
<td>71.90</td>
<td>71.48</td>
<td>64.52</td>
<td>73.21</td>
<td>72.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRD</td>
<td>75.48</td>
<td>71.46</td>
<td>73.48</td>
<td>75.51</td>
<td>73.94</td>
<td>69.73</td>
<td>75.65</td>
<td>76.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIST</td>
<td>75.35</td>
<td>71.68</td>
<td>73.86</td>
<td>75.79</td>
<td>73.86</td>
<td>69.17</td>
<td>76.08</td>
<td>75.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRRL</td>
<td>75.46</td>
<td>71.51</td>
<td>73.80</td>
<td>75.92</td>
<td>73.23</td>
<td>69.34</td>
<td>75.66</td>
<td>76.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SemCKD</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>76.23</td>
<td>74.43</td>
<td>69.61</td>
<td>77.62</td>
<td>76.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

UniADS (Ours) 76.32 71.93 74.20 76.65 75.12 70.79 79.73 76.87

Table 3: Results comparison of our UniADS with other methods (e.g., FitNets (Romero et al. 2015), AT (Zagoruyko and Komodakis 2017), SP (Tung and Mori 2019), CC (Peng et al. 2019), RKD (Park et al. 2019), CRD (Tian, Krishnan, and Isola 2020), DIST (Huang et al. 2022), SRRL (Jing Yang 2021) and SemCKD (Liu, Liu, and Huang 2022)). We report Top-1 “mean (std)” accuracies (%) for UniADS over 3 runs under the same CRD (Tian, Krishnan, and Isola 2020) training settings.

Ablation Study

In this section, we isolate the impact of each component of our UniADS and study other possible variants. All experiments are conducted on CIFAR-100 dataset. For the experiments of each setting, we run our method 3 times and report top-1 “mean (std)” accuracies.

**Hyperparameter Importance Analysis.** To analyze the impact of different search objects in the search space on the total distillation results, we used the hyperparameter optimization (HPO) tool to analyze the importance of each element. We added optimization parameters to the search (e.g., batch size, learning rate) in Figure 3. Through the different important analyses, it is discovered that (1) The network’s depth and width significantly impact the search results, particularly the network width, which is consistent with the search space design. (2) The weights assigned to transformations and feature losses for the knowledge output from the network have a considerable impact, aligning with the characteristics of the knowledge itself. (3) Some additional optimization parameters exhibit minimal influence. These observations provide valuable insights for optimizing knowledge distillation and
Table 4: The accuracy (%) of ResNet18 on ImageNet-1k with various teachers. Results of other KD methods refer to the papers of CRD (Tian, Krishnan, and Isola 2020), ONE (Zhu, Gong et al. 2018), DML (Zhang et al. 2018) and ESKD (Cho and Hariharan 2019). ATKD $A_{R34}$ (Mirzadeh et al. 2020) denotes ResNet34 used as the assistant teacher. N/A means no available results.

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Student</th>
<th>Acc.</th>
<th>Teacher</th>
<th>Student</th>
<th>KD</th>
<th>ESKD</th>
<th>ATKD $A_{R18}$</th>
<th>ONE</th>
<th>DML</th>
<th>DisWOT</th>
<th>UniADS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet34</td>
<td>ResNet18</td>
<td>Top-1</td>
<td>73.40</td>
<td>Top-1</td>
<td>69.75</td>
<td>70.66</td>
<td>70.89</td>
<td>70.78</td>
<td>70.55</td>
<td>71.03</td>
<td>72.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Top-5</td>
<td>91.42</td>
<td>Top-5</td>
<td>89.07</td>
<td>89.88</td>
<td>90.06</td>
<td>89.99</td>
<td>89.59</td>
<td>90.28</td>
<td>90.38</td>
</tr>
<tr>
<td>ResNet50</td>
<td>ResNet18</td>
<td>Top-1</td>
<td>76.16</td>
<td>Top-1</td>
<td>69.75</td>
<td>70.68</td>
<td>70.65</td>
<td>70.85</td>
<td>70.65</td>
<td>70.70</td>
<td>70.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Top-5</td>
<td>92.86</td>
<td>Top-5</td>
<td>89.07</td>
<td>N/A</td>
<td>N/A</td>
<td>70.85</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Figure 4: Optimal search curves of random search (Left), NSGA-II genetic algorithm with the Successive Halving algorithm (Middle), and NSGA-II genetic algorithm (Right) for 100 trials in ResNet experiments with ResNet110 as the teacher model on CIFAR-100.

**Conclusion**

In this paper, we present UniADS, an automated search framework for optimizing architectural dimensions and distiller settings in knowledge distillation. UniADS offers a comprehensive and efficient approach by exploring a wide range of design choices, including various architectural dimensions, knowledge transformations, distance functions, loss weights, and other crucial settings. To achieve this, UniADS utilizes the NSGA-II multi-objective optimization algorithm, which balances accuracy and performance trade-offs as conflicting objectives. By evaluating different configurations and applying crossover and mutation operations, NSGA-II generates new candidate solutions iteratively, aiming to improve the overall performance of the student model while maintaining a diverse set of high-quality solutions. Additionally, UniADS incorporates the Successive Halving algorithm, an acceleration strategy that progressively eliminates underperforming configurations, effectively reducing the computational burden and speeding up the distiller search process. Extensive experiments and comparative studies on two benchmark datasets provide evidence of the effectiveness and universality of the UniADS framework across various models. Notably, UniADS achieves outstanding accuracy on CIFAR-100 and ImageNet datasets, demonstrating its potential to contribute new ideas and methods to advance knowledge distillation methods. We hope that this research result is expected to provide new ideas and methods for the development of knowledge distillation techniques and better solutions for the compression and acceleration of deep learning models.

**Acknowledgments**

This work is supported by the National Natural Science Foundation of China under Grants No. 62206128 and No. 62302223.

**References**


Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei,
Dong, P.; Li, L.; and Wei, Z. 2023. Diswot: Student architecture search for distillation without training. In CVPR.