UniADS: Universal Architecture-Distiller Search for Distillation Gap

Liming Lu[1*], Zhenghan Chen[2*], Xiaoyu Lu[1†], Yihang Rao[1], Lujun Li[3], Shuchao Pang[1,4†]

[1]School of Cyber Science and Engineering, Nanjing University of Science and Technology
[2]Peking University
[3]HKUST
[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 overconfidence 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.

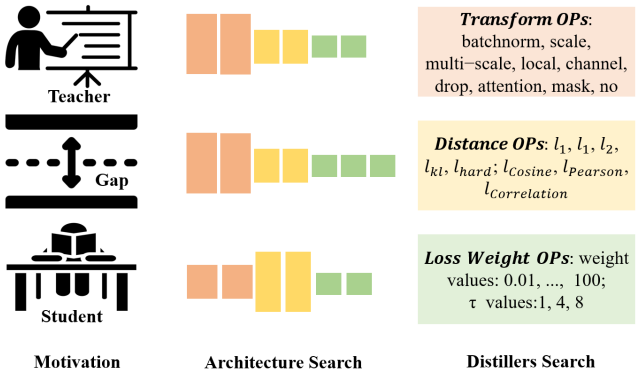

Figure 1: Brief overview of our UniADS. UniADS co-search student architecture in depth & width and distillation settings, including transform, distance, and loss weights to alleviate distillation gaps.

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-

*These authors contributed equally.
†Corresponding author.

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 mimicking, 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

Architecture	Depth,Width	Depth-{1,2,3} values:1,3,5,7; Width-ratio: 0.5,1,1.5,2
Distiller	Transform	$batchnorm, scale, multi-scale, local, batch, channel, drop, satt, natt, catt, mask, no$
	Distance	no-norm loss: smooth ℓ_1, ℓ_1, ℓ_2, ℓ_{KL}, ℓ_{hard}; norm loss: ℓ_{Cosine}, $\ell_{Pearson}$, $\ell_{Correlation}$
	Weight	weight values: 0.01,...,100; τ values:1,4,8,16

Table 1: Specific operations in UniADS.

changes or manual KD design efforts. UniADS transcends the focus of meta-KD (Deng et al. 2022; Liu et al. 2022) methods, which mainly fine-tune hyperparameters amidst complex optimization landscapes. It ventures into the uncharted territory of distiller design, revolutionizing KD research and practice. By simultaneously navigating hyperparameters and distillation strategies, UniADS provides a holistic approach to knowledge distillation, enhancing the efficiency and effectiveness of knowledge transfer.

Distillation-aware architecture search. To address the gap in teacher-student architectures, distillation-aware architecture search (DAS) has emerged as an essential task. The aim of DAS is to find an optimal student model architecture that can effectively utilize the teacher model's knowledge (Wei et al. 2024; Hu et al. 2021; Dong et al. 2022; Chen et al. 2022; Dong, Li, and Wei 2023; Dong et al. 2023; Zimian Wei et al. 2024). One of the initial methods (Liu et al. 2020) for DAS used a traditional reinforcement learning (RL) search techniques. This method used the KD loss as a feedback signal to guide the search process. However, it required significant computational resources and time due to the RL-based search. To address these limitations, subsequent work (Gu and Tresp 2020) introduced more efficient, gradient-based methods for DAS. These new methods utilize gradient information to refine and optimize the student model's architecture during the search. Despite their efficiency, these gradient-based approaches can be challenging due to issues like non-differentiability and the high dimensionality of the search space. A recent innovation in this area is the DisWOT (Dong, Li, and Wei 2023) framework. DisWOT is a gradient-free DAS framework that reduces the computational load by eliminating the need for repeated training iterations. It uses a pre-trained teacher model and a dataset to inform the search process without further training. However, many existing DAS methods do not adequately consider the distiller settings, which include the strategies and configurations that enable effective knowledge transfer from the teacher to the student model. Overlooking these settings can lead to suboptimal distillation results and maintain a significant gap between the teacher and student models. In contrast, our proposed framework extends beyond the conventional methods by including hyper-parameter search that encompasses transformations, distance functions, loss weights, and architectures. By thoroughly exploring this complex search space, our UniADS framework is designed to fine-tune the distillation process and bridge the gap between the teacher and student models effectively. It offers a more systematic and automated approach to distiller design, leading to more efficient knowledge transfer and improved distillation outcomes.

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:

$$\mathcal{L}_{KD} = \tau^2 \times \mathcal{W} \times \mathcal{D}\big(\mathcal{T}(\mathcal{A}_S)/\tau, \mathcal{T}(\mathcal{A}_T)/\tau\big), \quad (1)$$

where $\mathcal{W}$ is the loss weights factor, τ is the temperature factor, $\mathcal{T}$ is transformations, $\mathcal{D}(\cdot,\cdot)$ is distance function measuring the knowledge difference. $\mathcal{A}$ 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 ℓ_1, ℓ_1, ℓ_2, ℓ_{KL}, ℓ_{hard}, ℓ_{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 α^* from search space $\mathcal{A}$, as:

$$\alpha^* = \arg\min_{\alpha \in \mathcal{A}} (\mathcal{L}_{CE}(f(\mathcal{A}_S), Y) + \|Param.(\mathcal{A}_S) - \mathcal{C}\|), \quad (2)$$

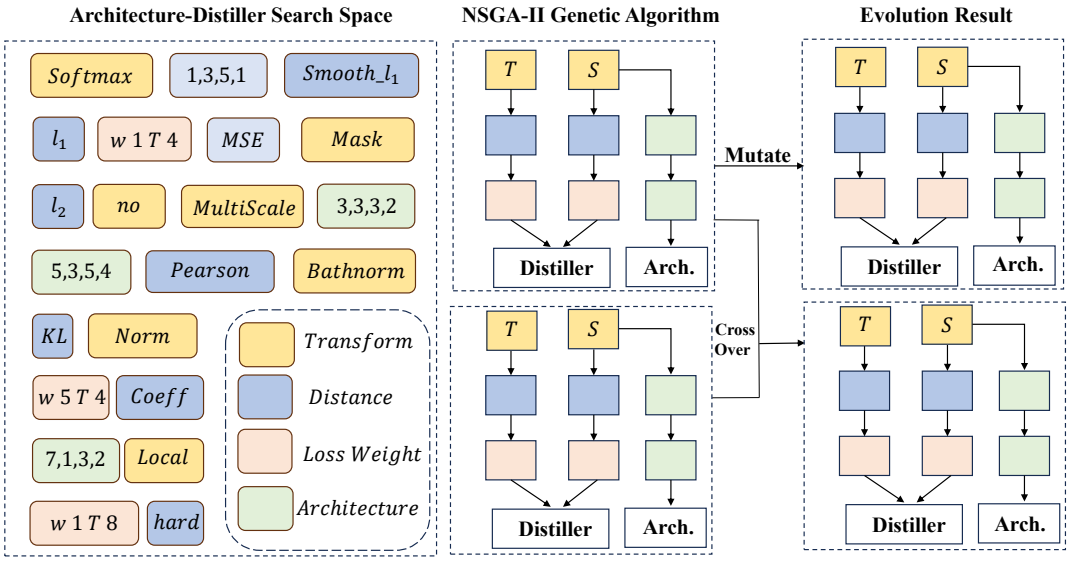

Figure 2: The overall illustration of our UniADS. Our architecture-distiller search space includes options for student depth & width, transformations, distance functions, and loss weights. From these options, we construct our architecture-distiller search and use the NSGA-II genetic algorithm with successive halving to search for the best configurations.

where $\mathcal{L}_{CE}$ is the regular cross-entropy objective with labels Y. $\mathcal{C}$ 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 $\mathcal{T}$: 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 $\mathcal{D}(\cdot, \cdot)$: The most frequently occurring metrics are "cos", "kl" and "l2". This suggests that some

Teacher	KD	DAS	UniDAS	$\mathcal{T}_f$	$\mathcal{T}_e$	$\mathcal{T}_p$	$\mathcal{D}(\cdot,\cdot)$	τ	Depth-1	Depth-2	Depth-3	Width-ratio
ResNet56	70.76	71.01	72.45	batchnorm	scale_r2	softmax_N	kl	4	7	1	3	1
ResNet110	72.18	72.58	74.89	mask	mask	softmax_C	l2	16	7	3	5	1
WRN-40-2	73.36	73.94	75.33	tanh	scale_r1	scale	cor	8	3	3	5	2
ResNet32x4	74.42	74.86	76.25	drop	natt	bmm	kl	8	7	3	5	1.5

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.

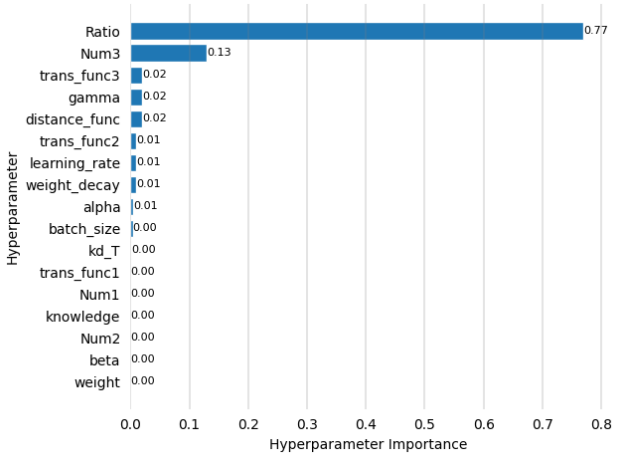

Figure 3: Hyperparameter importance of our UniADS search for ResNet experiments on CIFAR-100. Ratio, Num3, trans_func3, gamma denote the model width-ratio, model depth configuration, knowledge transform in logits and loss weights.

common distance functions can also provide promising distillation results.

- Loss weight $\mathcal{W}$: The table includes different loss weights assigned to components in the distillation loss function. The values 16, 5, 1, and 0.5 frequently appear as loss weights. These weights determine the relative importance of different components in the overall distillation loss. For instance, a higher weight indicates a greater emphasis on a particular component during the distillation process.
- Temperature τ: Relatively large values such as 16 and 8 are often chosen, which may help to alleviate teachers' overconfidence.
- Regarding architecture, wider and deeper networks tend to have good performance. Moreover, the width of the network and the number of cells in the deeper layers of the network have larger importance.

The examination of the findings in Table 2 yields valuable insights that can inform the design of knowledge distillation techniques. These insights offer practical guidance for effectively implementing KD methods.

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

Teacher Student	WRN-40-2 WRN-16-2	ResNet110 ResNet20	ResNet110 ResNet32	ResNet32x4 ResNet8x4	VGG13 VGG8	VGG13 MobileNetV2	ResNet32x4 ShuffleNetV2	WRN-40-2 ShuffleNetV1
Tea_Acc	75.61	74.31	74.31	79.42	74.64	74.64	79.42	75.61
Stu_Acc	73.26	69.06	71.14	72.50	70.36	64.60	71.82	70.50
KD	74.92	70.67	73.08	73.33	72.98	67.37	74.45	74.83
FitNet	73.58	68.99	71.06	73.50	71.02	64.14	73.54	73.73
AT	74.08	70.22	72.31	73.44	71.43	59.40	72.73	73.32
SP	73.83	70.04	72.69	72.94	72.68	66.30	74.56	74.52
CC	73.56	69.48	71.48	72.97	70.71	64.86	71.29	71.38
RKD	73.35	69.25	71.82	71.90	71.48	64.52	73.21	72.21
CRD	75.48	71.46	73.48	75.51	73.94	69.73	75.65	76.05
DIST	75.35	71.68	73.86	75.79	73.86	69.17	76.08	75.85
SRRL	75.46	71.51	73.80	75.92	73.23	69.34	75.66	76.61
SemCKD	N/A	N/A	N/A	76.23	74.43	69.61	77.62	76.39
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.

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 state-of-the-art (SOTA) KD methods. We conduct multiple trials with randomly selected distillers in the same search space to evaluate the efficacy of our EA search. For teacher-student pairs with the same architectural style, UniADS outperforms the baselines by margins ranging from $1.12\% \sim 3.32\%$. Compared with CRD and other KDs, UniADS obtains consistent performance gains ($2.38\% \sim 6.52\%$). Besides strengths in the same architecture pairs, UniADS exhibits even stronger performance when dealing with different architectural styles, while other KD methods suffer from noticeable accuracy reductions. Specifically, UniADS outperforms the baseline by margins of $1.12\% \sim 5.28\%$ and the random search results by margins of $1.9\% \sim 2.3\%$, demonstrating the effectiveness of our design for different structures. Compared with other SOTA KD methods, our UniADS achieves $1.2\% \sim 1.5\%$ gains. These results show that UniADS can improve each student model with simple settings under different teacher-student pairs.

Experiments on ImageNet

Dataset. We also conduct experiments on the ImageNet dataset (ILSVRC12) (Deng et al. 2009), which is known as one of the most challenging image classification datasets. It contains about 1.2 million training images and 50 thousand validation images, each belonging to one of the 1,000 categories.

Implementation. Following CIFAR-100 trials, we employ similar EA settings on a subset of ImageNet for search acceleration. Then, we utilize the discovered distiller for the training of student models (*e.g.*, ResNet-18 (He et al. 2016)). All teacher-student networks are trained with an SGD optimizer for 100 training epochs. The batch size is set to 256, and the learning rate is initialized to 0.1 and decays by 0.1 at 30, 60, and 90 epochs.

Comparison results. Table 4 reports the results of our methods on ImageNet. UniADS and improve baseline models of ResNet18 by 1.59% and 1.83% absolute gains in top-1 accuracy, respectively, which validate the effectiveness of our designs on the large-scale dataset. In addition, UniADS surpasses other SOTA methods with clear gains, demonstrating its superiority in large-scale datasets. These findings substantiate the effectiveness of UniADS in distillation optimization with considerable benefits, establishing the versatility and potency of our framework. In summary, UniADS facilitates substantially improved predictive accuracy of student models on ImageNet, more complex domains while preserving superior performance.

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 importance 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

Teacher	Student	Acc.	Teacher	Student	KD	ESKD	ATKD$_{A_{R18}}$	ONE	DML	DisWOT	UniADS
ResNet34	ResNet18	Top-1	73.40	69.75	70.66	70.89	70.78	70.55	71.03	72.08	72.25
		Top-5	91.42	89.07	89.88	90.06	89.99	89.59	90.28	90.38	90.58
Teacher	Student	Acc.	Teacher	Student	KD	ATKD$_{A_{R18}}$	ATKD$_{A_{R34}}$	Seq. ESKD	ESKD	DisWOT	UniADS
ResNet50	ResNet18	Top-1	76.16	69.75	70.68	70.65	70.85	70.65	70.95	72.30	72.51
		Top-5	92.86	89.07	N/A	N/A	N/A	N/A	N/A	90.51	90.77

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.

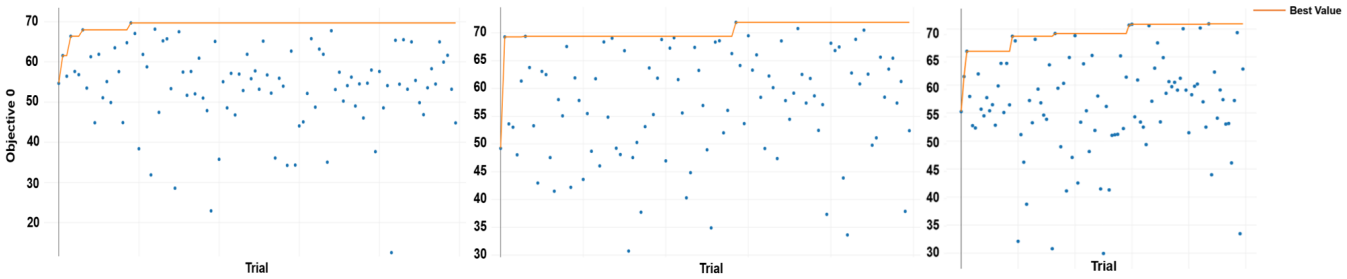

Figure 4: Optimal search curves of random search (Left), NSGA-II genetic algorithm with the Successive Halving algorithm (Middle), and NSGA-II genetic algorithm (Right) for 100 trials in ResNet experiments with ResNet110 as the teacher model on CIFAR-100.

guiding experiments in practical applications.

Search algorithm. We use the NSGA-II genetic for distiller search, which is gradient-free and flexible for non-convex optimization. Figure 4 demonstrates the effectiveness of the NSGA-II genetic algorithm compared to random search. The figure likely depicts a comparison of the convergence behavior and final search results between the two methods. It shows that NSGA-II achieves faster convergence, meaning it reaches good solutions more quickly and produces better final search results compared to random search. By integrating successive halving algorithms into the distiller search process, UniADS is able to achieve better search results with fewer trials. This implies that UniADS efficiently explores the search space, discarding unpromising configurations early on and allocating more resources to the most promising ones, leading to improved performance in distiller search.

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

Chen, K.; Yang, L.; Chen, Y.; Chen, K.; Xu, Y.; and Li, L. 2022. GP-NAS-ensemble: a model for the NAS Performance Prediction. In *CVPRW*.

Cho, J. H.; and Hariharan, B. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4794–4802.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei,

L. 2009. Imagenet: A large-scale hierarchical image database. 248–255.

Deng, X.; Sun, D.; Newsam, S.; and Wang, P. 2022. DistPro: Searching A Fast Knowledge Distillation Process via Meta Optimization.

Dong, P.; Li, L.; and Wei, Z. 2023. Diswot: Student architecture search for distillation without training. In *CVPR*.

Dong, P.; Li, L.; Wei, Z.; Niu, X.; Tian, Z.; and Pan, H. 2023. EMQ: Evolving Training-free Proxies for Automated Mixed Precision Quantization. *arXiv preprint arXiv:2307.10554*.

Dong, P.; Niu, X.; Li, L.; Xie, L.; Zou, W.; Ye, T.; Wei, Z.; and Pan, H. 2022. Prior-Guided One-shot Neural Architecture Search. *arXiv preprint arXiv:2206.13329*.

Gu, J.; and Tresp, V. 2020. Search for better students to learn distilled knowledge. *arXiv preprint arXiv:2001.11612*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

Hu, Y.; Wang, X.; Li, L.; and Gu, Q. 2021. Improving one-shot NAS with shrinking-and-expanding supernet. *Pattern Recognition*.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge Distillation from A Stronger Teacher. *arXiv preprint arXiv:2205.10536*.

Jing Yang, A. B. G. T., Brais Martinez. 2021. Knowledge distillation via softmax regression representation learning. In *ICLR2021*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Tech Report*.

Liu, J.; Liu, B.; Li, H.; and Liu, Y. 2022. Meta knowledge distillation. *arXiv preprint arXiv:2202.07940*.

Liu, Y.; Jia, X.; Tan, M.; Vemulapalli, R.; Zhu, Y.; Green, B.; and Wang, X. 2020. Search to Distill: Pearls are Everywhere but not the Eyes. In *CVPR*.

Liu, Z.; Liu, Y.; and Huang, C. 2022. Semi-Online Knowledge Distillation. *arXiv:2111.11747*.

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved knowledge distillation via teacher assistant. In *AAAI*.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *CVPR*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*.

Peng, B.; Jin, X.; Liu, J.; Li, D.; Wu, Y.; Liu, Y.; Zhou, S.; and Zhang, Z. 2019. Correlation Congruence for Knowledge Distillation. In *ICCV*.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. In *ICLR*.

Shu, C.; Liu, Y.; Gao, J.; Yan, Z.; and Shen, C. 2021. Channel-Wise Knowledge Distillation for Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5311–5320.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Representation Distillation. In *ICLR*.

Tung, F.; and Mori, G. 2019. Similarity-Preserving Knowledge Distillation. In *ICCV*.

Wei, Z.; Pan, H.; Li, L. L.; Lu, M.; Niu, X.; Dong, P.; and Li, D. 2024. TVT: Training-free Vision Transformer Search on Tiny Datasets. In *ICASSP)*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *BMVC*.

Zagoruyko, S.; and Komodakis, N. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *ICLR*.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4320–4328.

Zhou, H.; Song, L.; Chen, J.; Zhou, Y.; Wang, G.; Yuan, J.; and Zhang, Q. 2021. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective.

Zhu, X.; Gong, S.; et al. 2018. Knowledge distillation by on-the-fly native ensemble. *Advances in neural information processing systems*, 31.

Zimian Wei, Z.; Li, L. L.; Dong, P.; Hui, Z.; Li, A.; Lu, M.; Pan, H.; and Li, D. 2024. Auto-Prox: Training-Free Vision Transformer Architecture Search via Automatic Proxy Discovery. In *AAAI*.