LENAS: Learning-Based Neural Architecture Search and Ensemble for 3-D Radiotherapy Dose Prediction

Yi Lin, Graduate Student Member, IEEE, Yanfei Liu, Hao Chen, Senior Member, IEEE, Xin Yang, Member, IEEE, Kai Ma, Member, IEEE, Yefeng Zheng, Fellow, IEEE, and Kwang-Ting Cheng, Fellow, IEEE

Abstract—Radiation therapy treatment planning requires balancing the delivery of the target dose while sparing normal tissues, making it a complex process. To streamline the planning process and enhance its quality, there is a growing demand for knowledge-based planning (KBP). Ensemble learning has shown impressive power in various deep learning tasks, and it has great potential to improve the performance of KBP. However, the effectiveness of ensemble learning heavily depends on the diversity and individual accuracy of the base learners. Moreover, the complexity of model ensembles is a major concern, as it requires maintaining multiple models during inference, leading to increased computational cost and storage overhead. In this study, we propose a novel learning-based ensemble approach named LENAS, which integrates neural architecture search with knowledge distillation for 3-D radiotherapy dose prediction. Our approach starts by exhaustively searching each block from an enormous architecture space to identify multiple architectures that exhibit promising performance and significant diversity. To mitigate the complexity introduced by the model ensemble, we adopt the teacher–student paradigm, leveraging the diverse outputs from multiple learned networks as supervisory signals to guide the training of the student network. Furthermore, to preserve high-level semantic information, we design a hybrid loss to optimize the student network, enabling it to recover the knowledge embedded within the teacher networks. The proposed method has been evaluated on two public datasets: 1) OpenKBP and 2) AIMIS. Extensive experimental results demonstrate the effectiveness of our method and its superior performance to the state-of-the-art methods. Code: github.com/hust-linyi/LENAS.

Index Terms—Diversity, ensemble learning, knowledge distillation (KD), neural architecture search (NAS).

I. INTRODUCTION

RADATION therapy, chemotherapy, surgery, or their combination are widely utilized in clinical settings for cancer control [1]. Compared to conventional 3-D conformal therapy, modern treatment methods, such as intensity modulation radiation therapy and volumetric arc therapy, place greater emphasis on delivering the prescribed dosage to the planning target volume (PTV) while minimizing radiation exposure to organs at risk (OARs) [2]. Given the proximity of tumors to these critical structures, accurate tumor delineation is crucial to prevent radiation-induced toxicity [3]. The process of achieving a treatment plan with an optimal dose distribution involves meticulous iterations by a physicist, who adjusts various treatment planning parameters and weightings to balance clinical objectives [4]. However, this procedure is time consuming and prone to inter- and intra-observer variability, given the varying experience and skills of physicists [5].

Knowledge-based planning (KBP) presents an automatic solution that addresses the human effort required in traditional treatment planning by generating dose distributions, patient-specific dose-volume histograms (DVHs), and dose constraints for PTVs and OARs [6], [7], [8]. This approach serves as a reference to optimize the planning and control its quality, which effectively streamlines the treatment planning process. Recently, the field of KBP has witnessed significant advancements inspired by deep learning techniques [9]. Researchers have explored data-driven approaches to directly predict dose distributions. For instance, Nguyen et al. [10] utilized U-Net to predict dose distributions for prostate cancer, while Fan et al. [8] extended this approach by employing ResUNet for dose prediction in head-and-neck cancer. Kandalan et al. [11] investigated the generalizability of U-Net for prostate cancer dose prediction through transfer learning.
with limited input data. However, existing methods typically rely on U-Net and its variants [4], [10], which may not guarantee applicability across different physicists, diseases, and clinical settings.

Ensemble learning has been widely adopted in various deep learning tasks and has demonstrated impressive power [12]. These ensembles consist of multiple neural networks, with predictions combined through weighted averaging or voting during the test stage. While diversity is considered crucial for ensemble learning [13], existing methods often overlook it by employing a single-network architecture coupled with various training strategies or the combination of off-the-shelf architectures. The issue of limited diversity can be effectively addressed by leveraging neural architecture search (NAS) method, which can generate a large number of diverse architectures, driving a natural bias toward diversity of predictions, and in turn to afford the opportunity to integrate these networks to achieve improved results. However, several important research gaps are rarely explored. First, the relative importance of base learners’ performance and diversity in ensemble learning is not well understood. Second, striking a balance between ensemble performance and computational complexity is another crucial aspect that demands further investigation. Third, encouraging diversity within the search process of NAS deserves attention. To address the aforementioned challenges, we propose a learning-based ensemble approach with NAS for dose prediction, named LENAS. Our method adopts the teacher–student paradigm by leveraging a combination of diverse outputs from multiple automatically designed neural networks as a teacher model zoo to guide the target student network. The core of our LENAS includes two modules. First, instead of relying on off-the-shelf networks, we present a novel U-shape differentiable NAS framework, named U-NAS, which automatically and efficiently searches for neural architectures from enormous architecture configurations for the teacher ensembles. In addition, we design a diversity-encouraging loss to ensure both the high performance and diversity of the searched architectures. Second, to reduce the computational costs in the inference phase and meanwhile ensure high ensemble performance, we further present a knowledge distillation (KD) network with adversarial learning, named KDA-Net, which hierarchically transfers the distilled knowledge from the teacher networks to the student network.

The proposed methodology is evaluated using two publicly available datasets, the Open Knowledge-Based Planning (OpenKBP) dataset from the 2020 AAPM Grand Challenge and the AI Medical Innovation System (AIMIS) dataset from the 2021 Tencent AIMIS Challenge (Task 4). Our U-NAS framework achieves exceptional performance on the OpenKBP dataset, surpassing the performance of the champion of the AAPM challenge [14], [15], [16], [17]. Furthermore, our single LENAS model also outperformed state-of-the-art methods, securing first place in the AIMIS challenge. Our contributions are fivefold.

1) We introduce LENAS, a learning-based ensemble framework, which comprises the U-NAS framework for efficient and automatic architecture search, and KDA-Net for achieving a balance between efficiency and accuracy.

2) We propose a diversity-encouraging loss, which explicitly enhances the diversity among the searched models.

3) We design a hybrid loss that facilitates hierarchical transfer of knowledge from teacher ensembles to a single lightweight model, containing both hard (similarity loss) and soft (adversarial loss) constraints.

4) We provide in-depth analyses and empirical guidelines for generating and selecting base learners in ensemble.

5) Extensive experiments on two public datasets demonstrated the effectiveness of each module and the superior performance of our method over state-of-the-art methods.

II. RELATED WORK

A. Knowledge-Based Planning

KBP is realized by building an atlas-based repository or a mathematical model to predict the dosimetry such as dose distribution, entire DVH curve, and dose volume metrics utilizing previously optimized plans [18]. In atlas-based methods, manually designed geometric features are selected as metrics to measure the similarity between previous plans and a new plan. The previous parameters of the most similar plan are adopted as the initialization of the new plan optimization. Conversely, modeling methods use handcrafted features to regress and predict DVH of a new plan to guide the optimization processing [19]. The features include overlap volume histogram (OVH) [20], beams eye view (BEV) projections, overlap of regions of interest (ROIs), etc., which are applicable to both atlas-based and modeling methods.

Nevertheless, traditional KBP methods have limitations as they only predict 2-D or 1-D dosimetry metrics, failing to capture the complete spatial distribution of dosage. In recent years, researchers mainly focused on deep-learning-based KBP methods. Leveraging the powerful capabilities of CNNs to extract statistical and contextual features, these methods enable direct and highly accurate prediction of 3-D voxel-wise dose distributions. The inputs of deep-learning-based models usually are images (e.g., CT images and structure masks), and the architecture of models is mainly U-Net [4]. The two main directions for improving the performance of CNN-based dose prediction are: 1) designing different architectures, including modified U-Net [21], U-Res-Net [22], HD U-Net [4], and GAN-based [23] and 2) adding clinical parameters into inputs, such as isocenter [24], beam geometry information [25], and isodose lines and gradient information [26].

B. Ensemble Learning

Ensemble learning has shown impressive power in various deep learning tasks [12], a large amount of literature has provided theoretical and empirical justifications for its success, including Bayesian model averaging [27], enriching representations [28], and reducing stochastic optimization error [29]. These arguments reached a consensus that the individual learner in the ensembles should be accurate and diverse [13], [30]. To encourage the diversity of the
ensembles, the strategies for building ensembles typically include: 1) training the same network with various settings, such as bagging [31], random initializations [32], and different hyperparameters [33] (e.g., iteration, learning rate, and objective function) and 2) training different networks with the various architectures [34]. One of the most famous techniques is dropout [35], in which some of the neurons are dropped in each iteration, and the final model can be viewed as an ensemble composed of multiple different submodels.

For combining the predictions of each base model in an ensemble, the most prevalent method is majority voting [36] for classification and segmentation (which can be viewed as pixel-wise classification), and simple averaging [37] for the regression task. Despite their success, most existing ensemble methods do not explicitly balance the two important factors, i.e., the accuracy of individual learners and diversity among them. Different from these attempts, we reveal that the ensemble candidates produced by NAS could simultaneously guarantee the diversity and individual model’s accuracy, achieving superior ensemble performance.

C. Neural Architecture Search

NAS aims at searching for a desirable neural architecture from a large architecture collection. It has received increasing interest in various medical image analysis tasks, such as image classification [38], localization [39], segmentation [40], and reconstruction [41]. Much of the focus has been on the design of search space and search strategy. For example, Weng et al. [40] introduced NAS-Unet for 2-D medical image segmentation, which consists of different primitive operation sets for down- and up-sampling cells. Zhu et al. [42] proposed V-NAS for volumetric medical image segmentation that built a search space, including 2-D, 3-D, or pseudo-3D (P3D) convolutions. As for the search strategy, existing research can be categorized into three classes: 1) the evolutionary algorithm [43]; 2) reinforcement learning [44]; and 3) gradient-based differentiable methods [45].

In addition to searching for the single best model, recent works [46], [47], [48] proposed to search diverse architectures to form stronger and more robust ensembles. For instance, NEAS [49] and SAEP [47] encourage the disagreement between the base classifiers for ensemble pruning. NES [46] and NESS [48] optimize the sampled subarchitecture ensembles based on the evolutionary or sampling algorithm. Our method is different from the methods above in three ways. First, the ensemble pruning methods [47], [50] attempt to reduce the search space to facilitate the search process, while our method directly distills large ensembles into a single network. Second, the neural ensemble search methods [46], [48] aim to simultaneously search and retrain the ensembles, which is beyond the affordable computational capacity for 3-D medical images, where the input size is about 2.6k times larger than the CIFAR-10 dataset (i.e., 9 × 963 versus 3 × 322). Third, most existing methods [46], [49] focus on the classification task, using the disagreement of base classifiers as the diversity indicator, which is not applicable for regression tasks such as dose prediction.

III. METHODS

The framework of the proposed LENAS is shown in Fig. 1. It consists of two main components: 1) the U-NAS pipeline, a differentiable NAS for automatic architecture selection and 2) KDA-Net, which transfers the knowledge from the U-NAS ensembles to a lightweight network through adversarial learning. This enhances the models’ ability to extract meaningful features. On the other hand, KDA-Net plays a crucial role in reducing inference time while maintaining competitive accuracy. By combining these two components, U-NAS and KDA-Net, the LENAS framework achieves the dual objectives of enhancing feature extraction through diverse ensembles.
TABLE I
OPERATION SET USED FOR SEARCHING CELLS

| NormOps | DownOps | UpOps | ConnectOps | pre | post |
|----------|---------|-------|------------|-----|------|
| identity | avg_pool | up_se_conv | identity conv | conv | conv |
| se_conv | max_pool | up_dep_conv | no connection | connect | conv |
| dil_conv | down_se_conv | up_conv | dilated conv | down_dil_conv | up_dil_conv |
| dep_conv | down_dil_conv | interpolate | down_conv | 

and optimizing inference efficiency without compromising accuracy. In the following sections, we provide a detailed introduction to each component.

A. U-NAS

As shown in Fig. 1, the proposed U-NAS follows the autoencoder [17] structure with four down cells (DCs) and four up cells (UCs). Each individual cell is trained within an extensive search space comprising approximately 40,000 architecture configurations. In the following, we first introduce the search space and then describe the training strategy for joint optimization of the architecture and its weights.

**Search Space:** The yellow and red blocks in Fig. 1 show the network topologies of DC and UC, respectively, which include several fundamental computing units called hybrid modules (HMs). Each HM is a combination of different operations with four types: 1) normal (N); 2) downward (D); 3) upward (U); and 4) connect (C). These correspond to distinct operation groups within the search space. As shown in Table I, we include the following operations in the search space: convolution (conv), squeeze-and-excitation convolution (se_conv), dilated convolution (dil_conv), depthwise-separable convolution (dep_conv), max pooling (max_pool), average pooling (avg_pool), trilinear interpolation (interpolate), and residual connection (identity).

The prefix “down” means the stride of the convolution operation is two, while the prefix “up” indicates the transposed convolution, which doubles the image resolution. For the three columns of Table I, we use $3 \times 3 \times 3$ kernels for all convolution operations in the Conv-IN-ReLU order. In addition, a $3 \times 3 \times 3$ convolution (pre) and a $1 \times 1 \times 1$ convolution (post) are applied to adjust the number of channels.

**Training Strategy:** The training strategy of U-NAS contains two stages: 1) the search process and 2) the retraining process. In the search process, U-NAS is learned in a differentiable way [45], which optimizes a super network consisting of HMs with mixed operations. As Fig. 1 shows, for each operation $O_i$ in total $N$ operations $O$, the weight of each operation is determined by the parameter $\alpha_i \in \alpha$, whose softmax transformation $\tilde{\alpha}_i = \exp(\alpha_i)/\sum_{j=1}^{N} \exp(\alpha_j)$ represents how much $O_i$ contributes to the HM. Then, the architecture parameters $\alpha$ and the network weights $\omega$ are learned by the mixed operations alternately. To explore different local optima, the search process is repeated multiple times with different initializations, leading to the discovery of diverse architectures.

Table I: Operation Set Used for Searching Cells

| NormOps | DownOps | UpOps | ConnectOps | pre | post |
|----------|---------|-------|------------|-----|------|
| identity | avg_pool | up_se_conv | identity conv | conv | conv |
| se_conv | max_pool | up_dep_conv | no connection | connect | conv |
| dil_conv | down_se_conv | up_conv | dilated conv | down_dil_conv | up_dil_conv |
| dep_conv | down_dil_conv | interpolate | down_conv | |

Once the search process is completed, each HM retains only the most probable operation based on the parameter $\alpha$. Then, the DCs and UCs are replaced with the best-learned structures, and the network is retrained on the training dataset $D_{train}$. The detailed training strategy of U-NAS is outlined in Algorithm 1. During both the search and retraining processes, the difference between the predicted dose $\hat{y}$ and the target dose $y$ is measured using the $L_1$ norm

$$L_{dose} = \| y - \hat{y} \|_1. \quad (1)$$

**Diversity-Encouraging Loss:** The diversity among the models obtained by U-NAS can be potentially achieved via different initializations. However, in many settings, the independent search process could converge to similar local optima as the same search goal is exploited. To optimize for diversity directly in the architecture search process, we propose a diversity-encouraging loss to encourage different predictions between the learned model with the best model.

In the search process of U-NAS, the primary objective is to achieve high accuracy for the learned model while also encouraging differences in predictions between the learned model and the best model. Consequently, during the training stage of U-NAS, the final loss function $L_{nas}$ is formulated by combining the dose error loss $L_{dose}$ [in (1)] and the diversity-encouraging loss $L_{div}$ as follows:

$$L_{nas} = L_{dose} + L_{div}$$

$$= \| y - \hat{y} \|_1 + \eta \max \left(0, m - \frac{\| y - \hat{y} \|_1}{(\| y \|_1 + \| \hat{y} \|_1)/2} \right) \quad (2)$$

where $\| \cdot \|_1$ is the voxel-wise $l_1$ norm; $y$, $\hat{y}$, and $\hat{y}$ denote the ground-truth, prediction result of the training model and the best model, respectively; $m$ is the margin (empirically set to 0.2) used to reduce the correlation between $\hat{y}$ and $\hat{y}$ while avoiding the outliers; and $\eta$ is a weighting hyperparameter to balance the two loss terms (empirically set to 1).

B. KDA-Net

The proposed KDA-Net performs KD from the U-NAS ensembles to a single-target network with adversarial learning. As shown in Fig. 1, we use a single U-Net network as the student and the average of multiple U-NAS predictions as the

---

1The model with the best performance in multiple optimized architectures.
teacher ensemble. For all $K = 8$ blocks (four D blocks and four C blocks) of the network, we apply the similarity loss on the intermediate output [51] between the teacher ensembles and the student based on the squared Euclidean distance:

$$L_{\text{sim}} = \sum_{k=1}^{8} \left\| \frac{1}{M} \sum_{i=1}^{M} (\hat{I}_k^T - \hat{I}_k^S) \right\|_2^2$$

(3)

where $\hat{I}_k^T$ and $\hat{I}_k^S$ denote the intermediate output of the $k$th block of the $T$th teacher network and student network, respectively, and $M$ denotes the number of teacher networks.

To enhance the KD process, we incorporate adversarial learning, which promotes the generation of similar features by both the student and the teachers. This is achieved by introducing a discriminator $D$ for each block. The discriminator is responsible for distinguishing between the outputs of the teachers and the student. By doing so, the student is encouraged to produce outputs that closely resemble those of the teachers. The adversarial loss is defined as

$$L_{\text{adv}} = \sum_{k=1}^{8} \mathbb{E}_{I_k \sim P_T} \log D_k(I_k) + \sum_{k=1}^{8} \mathbb{E}_{I_k \sim P_S} \log(1 - D_k(I_k))$$

(4)

where $I_k \sim P_T$ and $I_k \sim P_S$ denote the outputs from the $k$th block of the teacher ensembles and the student network, respectively. Based on the above definition, we incorporate the dose loss in (1), the similarity loss in (3), and the adversarial loss in (4) into our KDA-Net loss function

$$L_{\text{KDA}} = L_{\text{dose}} + \lambda_1 L_{\text{sim}} + \lambda_2 L_{\text{adv}}$$

(5)

where $\lambda_1$ and $\lambda_2$ are weighting hyperparameters, which are empirically set to 0.05 and 0.01, respectively, in our experiments.

IV. EXPERIMENTS

A. Datasets

In this study, we evaluate the proposed method using two public datasets: 1) the OpenKBP dataset and 2) the AIMIS dataset.

OpenKBP Dataset: The OpenKBP dataset from the 2020 AAPM Grand Challenge [14] is a public dataset consisting of 340 CT scans for the dose prediction task. The OpenKBP dataset includes scans of subjects being treated for head-and-neck cancer with radiation therapy. The data is partitioned into training ($n = 200$), validation ($n = 40$), and test ($n = 100$) sets. The ROIs used in this study include the body, seven OARs (namely, brainstem, spinal cord, right parotid, left parotid, larynx, esophagus, and mandible), and three PTVs with gross disease (PTV70), intermediate-risk target volumes (PTV63), and elective target volumes (PTV65).

AIMIS Dataset: The AIMIS dataset from the 2021 Tencent AIMIS Challenge. Each scan is of a patient being treated for lung cancer with stereotactic body radiation therapy (SBRT). The dataset is officially partitioned into 300 scans for training.

2Instead of $L_1$ loss in (1), we adopt $L_2$ loss to the deep supervision for a fast optimization.

3https://contest.taop.qq.com/channelDetail?id=108

B. Implementation and Evaluation Metrics

The preprocessing for the two datasets follows [52]. For normalization, the CT values are truncated to $[-1024$ HU, 1500 HU]. The following data augmentations are performed during training: horizontal and vertical flips, translation, and rotation around the $z$-axis. For each sample of the OpenKBP dataset, the OAR masks (seven channels) and the merged target (one channel) are concatenated with the CT scan (one channel) as a $9 \times 128 \times 128 \times 128$ tensor and fed into the dose prediction models. For the AIMIS dataset, the input consists of OAR masks (five channels), CT scan (one channel), target volume (two channels), and body (one channel).

For the U-NAS search process, we first train the super network for $8 \times 10^4$ iterations using an Adam optimizer with an initial learning rate of $3 \times 10^{-4}$, and a weight decay of $1 \times 10^{-4}$. After that, the architecture parameters $\alpha$ are determined from the super network on the validation set. We repeat the search process multiple times with different random seeds to obtain various architectures. Then, we retrain the searched models on the training set for $8 \times 10^4$ iterations with a learning rate of $3 \times 10^{-4}$. For KDA-Net, we train the student network for $6 \times 10^4$ iterations using an Adam optimizer with an initial learning rate of $1 \times 10^{-4}$ and weight decay of $1 \times 10^{-4}$.

We use the official evaluation codes to validate the proposed method. Specifically, for the OpenKBP dataset, the evaluation metrics include 1) dose error, which calculates the mean absolute error (MAE) between the dose prediction and its corresponding ground-truth plan and 2) DVH error, which calculates the absolute error of the DVH curves between the prediction and ground truth. According to [14], $\{D_{99}, D_{50}, D_{1}\}$ for the PTVs and $\{D_{0.01cc}, D_{\text{mean}}\}$ for the OARs are selected to measure the similarity of the DVH curves in this task. For the AIMIS dataset, the evaluation is performed by measuring dose error with the mean-squared error (MSE). In addition, we use a paired $t$-test to calculate the statistical significance of the results.

C. Experimental Results

1) Performance of U-NAS: We first compare the performance of our U-NAS model with four manually designed architectures on the OpenKBP validation set. For each manually designed architecture, we employ the same convolution operation choice in HM, including conv, se_conv, dil_conv, and dep_conv. We apply max_pool, interpolate, and no_connection operations for all the manually designed architectures. Table II shows a performance comparison of different models. Our U-NAS model outperforms all the manually designed networks on the body, seven OARs, and three PTVs. The single U-NAS model achieves an MAE of 2.580 and 1.736 in dose score and DVH score, respectively, outperforming the best manually designed network by 0.111 and 0.128 in dose error and DVH error, respectively. It
is interesting, in most cases, the ensemble of four models outperforms the corresponding individual models (for both the manually designed and NAS learned models), and the ensemble of the NAS models outperforms the ensemble of the manually designed models. Please refer to Section V for more discussions of ensemble learning.

2) Performance of KDA-Net: We compare the performance of a single U-Net with and without the proposed KDA module, including the dose distributions and DVH, on the OpenKBP validation set. Fig. 2(a) shows an example of the DVH curves from a patient of the validation set. The solid lines represent the DVH curves of the ground truth, while the dashed and dotted lines represent the DVHs extracted from the predicted dose of U-Net with and without KDA (i.e., train from scratch), respectively. For this example patient, U-Net with KDA exhibits a better agreement in predicting the dose to the PTVs. The predictions of OARs are more variable between the two methods. Fig. 2(b) shows the corresponding dose color contour for the same patient as in Fig. 2(a), which suggests that the single U-Net model with KDA is able to achieve better dosimetric congruence with the original plan on the PTV.

3) Comparison With the State-of-the-Art Methods: In Table III, we compare the proposed LENAS model with several state-of-the-art methods on the OpenKBP test set. The competing methods include 3-D FCN [15], V-Net [15], 3-D U-Net [17], 3-D ResUNet [53], and five top-ranking methods on the AAPM-2020 challenge leaderboard [14]. We thoroughly compare our LENAS model with existing methods using single model, cascade, and ensemble strategies. The cascade strategy is to sequentially combine two networks and produce the results in a coarse-to-fine fashion. For single model, our U-NAS achieves an MAE of 2.597 and 1.803 in dose score and DVH score, respectively, outperforming the best off-the-shelf method (ResUNet). Integrating the KDA (i.e., LENAS) further improve the performance to 2.565 and 1.737, respectively. For cascade models, our cascade U-NAS model achieves 2.434 and 1.496 MAE in dose score and DVH score, respectively, outperforming the cascade ResUNet which achieves 2.448 and 1.499. For five model ensembles, our U-NAS ensemble achieves 2.357 MAE and 14.326 MSE in dose score, and 1.465 MAE and 5.560 MSE in DVH.
score, outperforming the ensembles of the off-the-shelf models and the top-ranking solutions on the AAPM-2020 challenge leaderboard.

4) Generalization Evaluation: We further explore the generalizability of our method on the AIMIS dataset. Specifically, we apply the best architecture (single model) learned from the OpenKBP dataset to the AIMIS challenge. The evaluation results are calculated by the organizers of the challenge, and shown in Table IV. Our U-NAS method achieves first place in the AIMIS challenge in both the primary and final phases, outperforming the runner-up by 9.36% and 1.88%, respectively. Moreover, in Table V, we further compare our U-NAS model with two best performing off-the-shelf models, U-Net and ResUNet, with respect to different ROIs. A consistent trend can be observed that our U-NAS outperforms the off-the-shelf models and other top-ranking solutions on both the validation set and the test set of AIMIS.

V. DISCUSSION

In this section, we investigate the correlations between diversity and ensemble performance, and empirically provide insightful guidance for ensemble learning with NAS in the task of dose prediction.

A. Ensemble Many Is Better Than All

In ensemble learning, most methods brusquely integrate all the obtained models to obtain the final result. To explore the correlation between the number of ensembles and their corresponding performance, we follow [54] to systematically conduct the search process multiple times and select the top 20 models for the experiment. Then, we average the results one by one sequentially w.r.t. the individual performance. The results are shown in Fig. 3. Fig. 3(a) shows the dose score (MAE) of the 20 selected models, which ranges from 2.5806 (NAS_1) to 2.6979 (NAS_20) MAE. Fig. 3(b) shows the dose score of ensembles of the top-\( k \in [1, 20] \) models. It can be seen that the ensembles achieve the best performance with the top 14 models (2.3621 MAE) instead of all the 20 models (2.3693 MAE). Intuitively, the inclusion of models with unacceptable performance in the ensembles could hurt the final ensemble results. Thus, the next step is to explore the selection criteria for the members of the ensemble.

B. Performance Versus Diversity

Extensive literature [13] has shown that the ideal ensemble is one comprised of accurate networks that make errors on different parts of the input space. In other words, the
C. Comparison of Different Ensemble Strategies

The uniqueness of the proposed LENAS model is to exploit a diverse collection of network structures that drive a natural bias toward the diversity of predictions produced by individual networks. To assess the impact of LENAS on the ensemble, we compare the diversity and performance of our method with four ensembling strategies, including bagging, random initializations, different training iterations, and different off-the-shelf architectures, on the OpenKBP validation set, as shown in Fig. 5. For each strategy, we obtain four models randomly. Specifically, for the NAS models, we select the top four models from the aforementioned 20 models (NAS_1 to NAS_4). For bagging, we split the training set into four nonoverlapped subsets using different portions of the data (three subsets) to train the four models. For random initialization, we repeat the training procedures four times with different initialization seeds. For the different training epochs, we train a single network with $8 \times 10^4$ iterations, and pick the last four checkpoints with a gap of 2000 training iterations in between. Finally, for the off-the-shelf architectures, we select the four most popular architectures in 3-D medical image analysis, including FCN, VNet, U-Net, and ResUNet.

The diversities of the four ensembling strategies are illustrated in Fig. 5(a). The diversity of the NAS models is 0.0326 RMAE with a standard deviation of 0.0009, greater than three of the other strategies, namely, cross-validation, random initialization and iterations, and comparable to off-the-shelf architectures, which achieve a diversity of 0.0338 $\pm$ 0.0013. Furthermore, the results presented in Fig. 5(b) demonstrate that the NAS models outperform other strategies by a significant margin in terms of the mean and standard deviation of the dose score, with a value of 2.587 $\pm$ 0.0082. Although the diversity of fourfold cross-validation is close to the NAS models, the individual models’ performance suffers from the limited representations of the subsets of training data, leading to a lower ensemble performance. Similar trends are observed in off-the-shelf models. Finally, the performance of the ensembles, as measured by the dose score in terms of MAE, is depicted in Fig. 6. It is evident that the ensembles of NAS models achieve the best performance, with an MAE of 2.392, surpassing other ensemble strategies. This result emphasizes the superiority of producing and ensembling different network architectures as opposed to creating an ensemble consisting solely of duplicates of a single-network architecture with different model parameters.
on two public datasets demonstrated the effectiveness and superior performance of our method compared to state-of-the-arts.

The LENAS method offers several notable advantages. First, the U-NAS framework enables automatic NAS, allowing for the efficient exploration of a vast array of architecture configurations. This search process results in the creation of a diverse teacher network zoo, which in turn facilitates the generation of a robust ensemble of learners. By leveraging this ensemble method, we are able to enhance the overall performance of the system. Second, the KDA-Net enables a hierarchical transfer of distilled knowledge from the teacher networks to the student network. This transfer mechanism not only reduces inference time but also ensures that the student network maintains a high level of accuracy, thus striking an optimal balance between efficiency and performance. Finally, the effectiveness of our approach has been thoroughly validated through rigorous experimentation on two public datasets. Notably, our method has achieved first place in both the AAPM and AIMIS challenges. This outstanding performance serves as compelling evidence of the promising potential and efficacy of our proposed LENAS method.

We would like to point out several limitations of our work. First, the NAS ensembles require multiple rounds of searching-retraining, which is very time consuming in the training phase. Second, a few failure models may be generated by NAS. This situation is also common in the gradient-based NAS methods. Third, the diversity between learners in an ensemble is hard to formulate appropriately, which could be task-specific and vary for different outputs (e.g., classification, segmentation, and regression). Future studies will be focused on: 1) a more specific model selection criterion for the best ensemble strategies; 2) a computational-efficient training strategy for multiple ensemble learners; and 3) an optimization method from dose prediction map to the radiotherapy treatment plan.

VI. CONCLUSION AND FUTURE WORK

In this article, we proposed a learning-based ensemble approach, named LENAS, for 3-D radiotherapy dose prediction. The two key components of LENAS are 1) a U-NAS framework which automatically searches neural architectures from numerous architecture configurations to form a teacher network zoo and 2) a KDA-Net which hierarchically transfers the knowledge distilled from the teacher networks to the student network to reduce the inference time while maintaining competitive accuracy. We conducted comprehensive experiments to investigate the impact of diversity in ensemble learning, and derived several empirical guidelines for producing and ensembling base learners in consideration of individual accuracy and diversity. Extensive experiments

We follow [45] to search single architecture of DC and UC in each model for facilitating the optimization of the search process.

REFERENCES

[1] H. Park, L. M. Sholl, H. Hatabu, M. M. Awad, and M. Nishino, “Imaging of precision therapy for lung cancer: Current state of the art,” Radiology, vol. 293, no. 1, p. 15, 2019.
[2] R. Atun et al., “Expanding global access to radiotherapy,” Lancet Oncol., vol. 16, no. 10, pp. 1153–1186, Sep. 2015.
[3] L. Lin et al., “Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma,” Radiology, vol. 291, no. 3, pp. 677–686, 2019.
[4] D. Nguyen et al., “3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture,” Phys. Med. Biol., vol. 64, no. 6, 2019, Art.no. 065020.
[5] B. Ye, Q. Tang, J. Yao, and W. Gao, “Collision-free path planning and delivery sequence optimization in noncoplanar radiation therapy,” IEEE Trans. Cybern., vol. 49, no. 1, pp. 42–55, Jan. 2019.
[6] G. Zhou et al., “All-in-one online radiotherapy for nasopharyngeal carcinoma: Preliminary results of treatment time, contouring accuracy, treatment plan quality and patient compliance,” Int. J. Radiat. Oncol., Biol., Phys., vol. 117, no. 2, pp. e636–e637, 2023.
[7] L. Yu et al., “First implementation of a one-stop solution of radiotherapy with full-workflow automation based on CT-linac combination,” Med. Phys., vol. 50, no. 3, pp. S117–S126, 2023.
[8] J. Fan, J. Wang, Z. Chen, C. Hu, Z. Zhang, and W. Hu, “Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique,” Med. Phys., vol. 46, no. 1, pp. 570–381, 2019.
[9] E. Huynh et al., “Artificial intelligence in radiation oncology,” Nat. Rev. Clin. Oncol., vol. 17, no. 12, pp. 771–781, 2020.

[10] D. Nguyen et al., “A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning,” Sci. Rep., vol. 9, no. 1, p. 1076, 2019.

[11] R. N. Kandalan et al., “Dose prediction with deep learning for prostate cancer radiation therapy: Model adaptation to different treatment planning practices,” Radiother. Oncol., vol. 153, pp. 226–235, Dec. 2020.

[12] B. Zolfaghari, L. Mirsadeghi, K. Bibak, and K. Kavousi, “Cancer prognosis and diagnosis methods based on ensemble learning,” ACM Comput. Surv., vol. 55, no. 12, pp. 1–34, 2023.

[13] Y. Bian and H. Chen, “When does diversity help generalization in classification ensembles?” IEEE Trans. Cybern., vol. 52, no. 9, pp. 9507–9505, Sept. 2022.

[14] A. Bahier et al., “RavenKBP: The open-access knowledge-based planning grand challenge,” 2021, arXiv:2011.14076.

[15] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 3341–3440.

[16] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Assist. Interv., 2015, pp. 234–241.

[17] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. 10th Int. Conf. Med. Image Comput. Comput. Assist. Interv., 2015, pp. 703–742.

[18] S. Momin et al., “Knowledge-based radiation treatment planning: A data-driven method survey,” J. Appl. Clin. Med. Phys., vol. 22, no. 8, pp. 16–44, 2021.

[19] X. Zhu, Y. Ge, T. Li, D. Thongphiew, F.-F. Yin, and Q. J. Wu, “A planning quality evaluation tool for prostate adaptive IMRT based on machine learning,” Med. Phys., vol. 38, no. 2, pp. 719–726, 2011.

[20] H. Li et al., “Explainable attention guided adversarial deep network for 3D radiotherapy dose distribution prediction,” Knowl.-Based Syst., vol. 241, Apr. 2022, Art. no. 108324.

[21] Y. Lin, D. Zhang, X. Fang, Y. Chen, K.-T. Cheng, and H. Chen, “Rethinking boundary detection in deep learning models for medical image segmentation,” in Proc. 28th Int. Conf. Inf. Process. Med. Imag., 2023, pp. 730–742.

[22] Z. Liu et al., “A deep learning method for prediction of three-dimensional dose distribution of helical tomotherapy,” Med. Phys., vol. 46, no. 5, pp. 1972–1983, 2019.

[23] D. Nguyen et al., “Incorporating human and learned domain knowledge into training deep neural networks: A differentiable dose-volume histogram-based and adversarial inspired framework for generating Pareto optimal dose distributions in radiation therapy,” Med. Phys., vol. 47, no. 3, pp. 837–849, 2020.

[24] S. Willems, W. Crijns, E. Sterpin, K. Haustermans, and F. Maes, “Feasibility of CT-only 3D dose prediction for VMAT prostate plans using deep learning,” in Proc. 1st Workshop Artif. Intell. Radiat. Therapy, 2019, pp. 10–17.

[25] L. Teng et al., “Beam-wise dose composition learning for head and neck cancer dose prediction in radiotherapy,” Med. Image Anal., vol. 92, Feb. 2024, Art. no. 103045.

[26] S. Tan et al., “Incorporating isodose lines and gradient information via multi-task learning for dose prediction in radiotherapy,” in Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Interv., 2021, pp. 753–763.

[27] Y. Yang, H. Lv, and N. Chen, “A survey on ensemble learning under the era of deep learning,” Artif. Intell. Rev., vol. 56, no. 6, pp. 5545–5589, 2023.

[28] Y. Yang, Y. Hu, X. Zhang, and S. Wang, “Two-stage selective ensemble of CNN via deep tree training for medical image classification,” IEEE Trans. Cybern., vol. 52, no. 9, pp. 9194–9207, Sep. 2022.

[29] Z.-H. Zhou, “Ensemble learning,” in Machine Learning. Singapore: Springer Singapore, 2021, pp. 181–210.

[30] D. Wool, T. Mu, A. M. Webb, H. W. Reeve, M. Lujan, and G. Brown, “A unified theory of diversity in ensemble learning,” J. Mach. Learn. Res., vol. 24, no. 359, pp. 1–49, 2023.

[31] N. Altman and M. Krzywinski, “Ensemble methods: Bagging and random forests,” Nat. Methods, vol. 14, no. 10, pp. 933–935, 2017.

[32] Y. Lin et al., “Nuclei segmentation with point annotations from pathology images via self-supervised learning and co-training,” Med. Image Anal., vol. 62, 2020.

[33] A. Mohammed and R. Kora, “A comprehensive review on ensemble deep learning: Opportunities and challenges,” J. King Saud Univ. Comput. Inf. Sci., vol. 35, no. 2, pp. 757–774, 2023.
Yanfei Liu received the M.S. degree from Hunan University, Changsha, China, in 2021. She was an Intern with the Jarvis Lab, Tencent, Shenzhen, China, from 2020 to 2021. She is currently a Research Cooperation Specialist with the Shenzhen United Imaging Research Institute of Innovative Medical Equipment, Shenzhen. Her research interests include medical image analysis and deep learning in radiation treatment planning.

Hao Chen (Senior Member, IEEE) received the Ph.D. degree from The Chinese University of Hong Kong (CUHK), Hong Kong, in 2017. He is an Assistant Professor with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong. He was a Postdoctoral Research Fellow with CUHK previously. He has 100+ publications in MICCAI, IEEE TRANSACTIONS ON MEDICAL IMAGING, Medical Image Analysis, CVPR, AAAI, Radiology, Lancet Digital Health, Nature Machine Intelligence, and JAMA. He also holds a dozen of patents in artificial intelligence (AI) and medical image analysis. His interests focus on developing trustworthy AI for healthcare.

Xin Yang (Member, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA, USA, in March 2013. She is currently a Professor with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China. Her research interests include medical image analysis, 3-D vision, monocular simultaneous localization, and mapping.

Kai Ma (Member, IEEE) received the Ph.D. degree from the University of Illinois at Chicago, Chicago, IL, USA, in 2014. He was with Siemens Medical Solution, Princeton, NJ, USA, for more than five years. He is currently a Principal Researcher with the Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China. His research interests include medical image analysis, deep learning, computer vision, and brain–computer interface.

Yefeng Zheng (Fellow, IEEE) received the B.E. and M.E. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 1998 and 2001, respectively, and the Ph.D. degree from the University of Maryland at College Park, College Park, MD, USA, in 2005. He was a Principal Key Expert with Siemens Healthineers, Princeton, NJ, USA, working on medical image analysis. He is currently the Director of the Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China. His research interests include medical image analysis, computer vision, deep learning, and natural language processing.

Kwang-Ting Cheng (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science from the University of California at Berkeley, Berkeley, CA, USA, in 1988. He holds the position of Vice-President for Research and Development with The Hong Kong University of Science and Technology, Hong Kong. Additionally, he has successfully transferred several of his inventions into commercial products. His contributions have garnered recognition from the academic community. He has an extensive publication record, with over 500 technical papers, coauthorship of five books, and 12 U.S. patents. His research interests primarily lie in computer vision, medical image analysis, and electronic design automation.

Dr. Cheng is a Fellow of the Hong Kong Academy of Engineering Sciences.