T-AutoML: Automated Machine Learning for Lesion Segmentation using Transformers in 3D Medical Imaging

Dong Yang Andriy Myronenko Xiaosong Wang Ziyue Xu Holger R. Roth Daguang Xu
NVIDIA

Abstract

Lesion segmentation in medical imaging has been an important topic in clinical research. Researchers have proposed various detection and segmentation algorithms to address this task. Recently, deep learning-based approaches have significantly improved the performance over conventional methods. However, most state-of-the-art deep learning methods require the manual design of multiple network components and training strategies. In this paper, we propose a new automated machine learning algorithm, ***T-AutoML****, which not only searches for the best neural architecture, but also finds the best combination of hyper-parameters and data augmentation strategies simultaneously. The proposed method utilizes the modern transformer model, which is introduced to adapt to the dynamic length of the search space embedding and can significantly improve the ability of the search. We validate T-AutoML on several large-scale public lesion segmentation data-sets and achieve state-of-the-art performance.*

1. Introduction

Nowadays, given the technological advances in algorithmic design (such as deep learning) and hardware platforms (such as GPU), medical image analysis has become a key step in disease understanding, clinical diagnosis, and treatment planning. The sizes, shapes, and appearances of lesions in medical images can vary greatly across different anatomical structures (shown in Fig. 1). These semantic features are closely related to the severity of the disease. At the same time, variations in the imaging patterns related to pathologies, scanning protocols, and medical devices introduce a computational challenge for automated algorithms [1]. Therefore, automated lesion segmentation is one of the most challenging tasks in automated analysis of medical images, such as 3D CT, 3D MRI, histopathology, etc. When facing with such challenges, most machine/deep learning models struggle to provide comprehensive solutions.

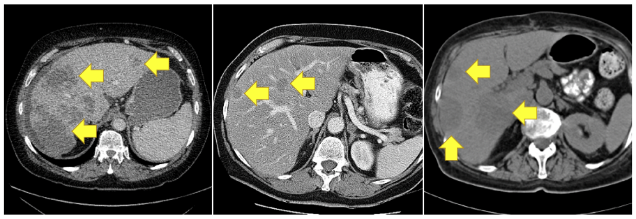

Figure 1. An example of intensity variations of liver lesions in 3D CT, which introduces substantial challenges for automated lesion segmentation. The yellow arrows indicate the target lesion locations. There are several large lesions in the left figure and small ones in the middle figure. The right figure presents an atypical CT contrast of liver and lesion regions after image windowing/normalization.

In recent years, deep convolutional neural networks have been widely applied for medical image analysis. For instance, one of the most influential works in medical image segmentation using deep neural networks is U-Net [2, 3], which has been found to be more effective and efficient than previous non-learning based methods. Furthermore, automated machine learning (AutoML) with neural networks has been explored for various applications, such as image recognition, semantic segmentation, object detection, natural image generation, etc. One of the most popular tasks in AutoML is neural architecture search (NAS) [4], which aims to design neural network architectures automatically without much human heuristics or assumptions. Besides the model weights, the model structure itself becomes fitted for the task after searching, and is even transferable to different applications [5]. Additional constraints, such as latency or parameter quantity of models, can be added as searching objectives to fit the models into different computing platforms. Several works [6] have achieved state-of-the-art (SOTA) performance with NAS in large-scale image recognition (*e.g.*, on ImageNet [7]). Other works applied similar techniques to optimize other components (*e.g.*, augmentations, loss functions) or hyper-parameters in conventional deep learning solutions. Although the recent development of AutoML has shown promising results in reducing the workload of algorithm design and improving model performance, most AutoML methods only optimize few components of the framework, *e.g.*, network or transform, and continue to rely on the human choice of other components, which can lead to sub-optimal solutions.

In this paper, we propose a new method to automatically estimate "almost" all components of a deep learning solution for lesion segmentation in 3D medical images. At first, a new search space for segmentation networks is introduced to enable flexible connection of global network structure beyond the U-shape design (encoder-decoder based models), which is commonly used in medical image segmentation [8]. Then, candidates of deep learning configurations (neural architecture, augmentations, hyper-parameters) are encoded to a 1D vector as the abstract representation. Next, a binary relation predictor is trained with the representative vectors of configurations and their corresponding validating metrics. More specifically, the predictor distinguishes between two input vectors to see if one could lead to better performance than the other. Given such predictors, all configurations of deep learning solutions can be sorted with a direct comparison. Finally, the search configurations are generated by sampling candidates from the candidate pool and picked by their predicted performance in the searched task of the lesion segmentation. Furthermore, the searched configurations can be transferred to similar tasks in different datasets and achieve reasonable performance.

The contributions of our proposed method can be summarized as follows.

1. We propose a new search space for lesion segmentation in 3D volumes, which has much more flexibility compared to the traditional U-shape architecture in medical imaging, and our architecture searching provides insights about the importance and fitness of various dense connections inside segmentation networks;

2. Our method reaches the state-of-the-art performance of lesion segmentation on two public datasets;

3. Both the search process and the configuration deployment of the proposed method is designed to be computationally efficient and effective;

4. To the best of our knowledge, this is the first time that a comprehensive AutoML (including both NAS and optimization of other related training hyper-parameters) is applied to medical image segmentation leveraging the capacity of transformer modules.

2. Related Work

Lesion segmentation: Automated lesion segmentation in medical imaging has been studied for decades. Some early studies involved handcrafted rules such as convergence index filter [9] and multi-scale hessian-based blob measurements [10]. Later, researchers proposed to adopt machine learning models to tackle this challenging task. The standard pipeline is to extract a series of handcrafted features, then distinguish lesion from the background by utilizing various classifiers, including the k-nearest neighbor [11], random forest [12, 13], and support vector machine [14]. The models can further leverage the advantage from ensembling with sufficient interpretation capacity. However, the reported model is usually very sensitive to imaging appearance with numerous false positives [13]. Recently, deep neural networks have been introduced to localize and segment lesions from multiple modalities, including MRI [15], ultrasound [16], and CT [17, 18], given several publicly available annotated datasets. The majority of these techniques take a network specifically designed for one previous task, and then do the training by pre-processing the data of the current task to accommodate the network design.

Efficient ConvNet models for medical imaging: Researchers have adopted the design concepts of U-Net ("U"-shape convolutional encoder-decoder architecture with skip connections), and further extended it into novel and effective neural network architectures [19, 20, 21, 22, 23, 24, 25, 26, 27]. Nowadays, 2D/3D U-Net is considered as the common baseline model for most segmentation tasks. Often state-of-the-art performance in several tasks is reached via using variants of U-Net with model ensembling [28]. However, one potential downside of U-Net is that once the U-shape models are trained properly, their prediction relies more on local context instead of the global structure because of the multi-level skip connections. Then densely connected U-Net (*e.g.*, U-Net++) was proposed to further improve the performance and generality of the segmentation models [29]. It gives us a hint that dense connection is helpful, but we have very limited intuition as to which connection plays a more important role and which connection can be pruned. Recently, researchers used neural architecture search to answer such questions to improve the performance of 3D segmentation networks further.

Automated machine/deep learning: Neural architecture search (NAS) is commonly used to design neural networks automatically with limited human heuristics to meet different user requirements (*e.g.*, light-weight models or minimal computation) [30]. Zoph and Le firstly introduced a novel framework to conduct neural architecture search using reinforcement learning for large-scale image recognition [4, 5, 31]. Specifically, they formulated the NAS problem as searching convolutional layers for repeated modules inside the entire network, which is capable of reducing the searching space and cost to a great extent since the same module structure is applied for several modules in the entire architecture. Their searched networks (after being trained) are able to achieve state-of-the-art performance on the ImageNet dataset [32]. Others followed a similar idea to adopt reinforcement learning (RL) to tune the hyper-parameters of training settings or to find optimal data augmentation solutions [33, 34]. Using newly searched neural networks, researchers have improved the performance of various computer vision applications (semantic segmentation, object detection, design of loss function, etc.) in some challenging datasets [35, 36, 37, 38, 39]. Meanwhile, researchers have explored other concepts of NAS, *i.e.*, "super-net" and one-shot NAS [40, 41, 42, 43, 44, 45, 46, 47, 48]. Another trend in NAS is predictor-based methods [49, 50, 51]. The predictors do not have performance gaps of searched architectures between searching and re-training. However, they require many trained models with validation accuracy as the ground truth to train a robust predictor model.

Recently, the related research has been widely developed in automated deep learning (AutoDL) to optimize almost every aspect of deep learning, including network architecture design [4, 52, 43, 6, 53, 53], data augmentation strategies [33, 34, 54], and loss functions [55], etc. With those automated components of deep learning, several computer vision applications successfully perform much better in various metrics (accuracy, latency, model size, etc.). However, previous works mainly focus on optimizing few components of a deep learning framework only. Our proposed framework aims to simultaneously search for the optimal combination of deep learning components, introducing greater flexibility to AutoDL.

3. T-AutoML

The success of deep learning comes from several aspects. One major advantage of deep learning is that neural networks are end-to-end trainable without the need for feature engineering. However, there is a new need for designing the best network architecture, which often determines the upper-bound performance of deep learning models. A good architecture enables effective gradient back-propagation during training and feature learning. To further promote the performance, data augmentation during training is utilized to increase model robustness and mitigate the gap between the domains of training, validation, and testing datasets. Last but not least, hyper-parameters in model training are also critical for fast convergence and decent accuracy.

Our proposed transformer-based AutoML (T-AutoML) method (see Sec. 3.3) trains a predictor to compare performance between different training configurations (neural architecture, data augmentation, and hyper-parameters). T-AutoML leveraging the advanced capacities of the transformer modules [56] to handle the input sequence with dynamic sequence, which is suitable for neural architecture space with a various number of blocks/layers. The proposed method is capable of covering (almost) all components in a conventional deep learning solution. The neural network architecture, data augmentation, and other related hyper-parameters can be fit into the method with proper encoding strategies. And the combination of encoding is the reference for the predictor to determine the optimal architecture and training configurations for the target tasks.

3.1. SpineNet Searching Space

U-shape neural networks for segmentation are very popular in modern deep learning since they are good at image-to-image translation [21, 2, 3, 19, 28]. The encoder reduces the feature maps gradually, and meanwhile increases the filter numbers. The decoder reverts back the spatial resolution gradually to match the input shape. Also, the skip connections between the encoder and decoder in the U-shape have been validated to be effective in capturing the low-level image details for the fine segmentation boundaries [57]. Although such a design is simple and straightforward, it produces promising results in many segmentation applications, and has become the go-to network in the field of medical image segmentation. However, there is no evidence that if the "encoder, decoder, and skip connection" design patterns are optimal for the task at hand.

Hence, we propose a new search space of neural architectures to further study the network connections for image segmentation. Our method is inspired by SpineNet [58], which was proposed for object detection in natural image processing. Under the framework, the feature maps at different spatial levels of the network can be connected with each other arbitrarily. At the same time, the order of operations to increase or decrease the feature maps' spatial size can be arranged randomly. Thus, the search space not only contains U-shape networks or densely connected networks [29] but also covers other network topologies with asymmetric structure (see Fig. 2). Therefore, the search space is much larger than previous ones in the segmentation-related NAS literature [59, 60, 61, 62, 63, 64].

Like other methods in NAS, the search space consists of several blocks with different operations. In our method, the candidates of different blocks are 3D residual blocks, 3D bottleneck blocks, and 3D axial-attention blocks [65]. The residual and bottleneck blocks are effective in avoiding vanishing gradient [66]. The axial-attention block was originally proposed to resolve the issue of weak long-range dependency in the 2D plane for segmentation tasks [65]. Our axial-attention is the extension from 2D to two 3D versions. The first one is conducted axis-wise attention along $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ axes sequentially. And the second one is conducted along $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{Z}$ axes sequentially. Since the axial plane ($\mathcal{X}$-$\mathcal{Y}$ plane) is commonly imaged in higher resolution and is more informative with low-level image details compared to $\mathcal{X}$-$\mathcal{Z}$ plane and $\mathcal{Y}$-$\mathcal{Z}$ planes, we let the axial attention block process the $\mathcal{X}$-$\mathcal{Y}$ plane first, and then conduct $\mathcal{Z}$ axis-wise attention.

In practice, we construct the architecture with $\mathcal{N}$ blocks

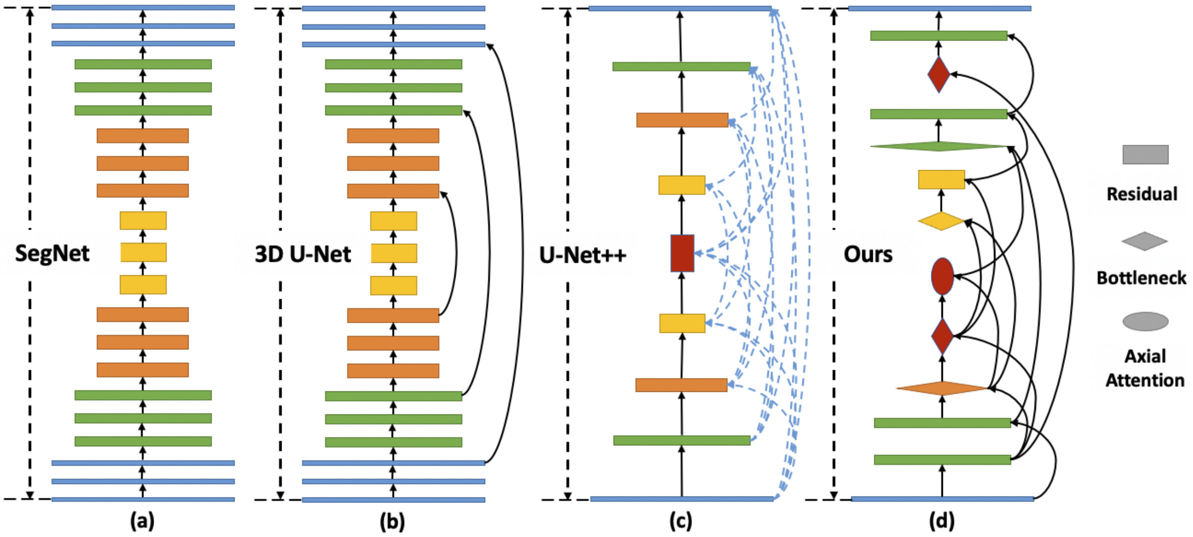

Figure 2. (a) SegNet [22]; (b) U-Net [3] (c) U-Net++ [29]; The connected dotted line shows connectivity, but the underlying operation is more complicated than a single convolution. (d) Our searched the architecture. Different colors indicate different spatial resolutions. Different block types are marked as different block shapes.

one-by-one. Whenever introducing a new block c to the architecture, its category and spatial resolution level of the block needs to be determined first. From the 3rd block, the block c_i would collect feature maps from two of all precedent blocks c_j and c_k, and combine them to a single feature map (c_2 would only receive the feature map from c_1. i, j, k are not necessarily adjacent). In order to combine layers from different spatial resolutions and match the resolution of the current block, necessary up-sampling and down-sampling are applied to the feature maps, respectively. We change the spatial resolution of combined feature maps to the target spatial resolution following the fashion from [58]. Then the combined feature maps would be converted to c_i's spatial resolution level with necessary up-sampling/down-sampling. Lastly, the $\mathcal{N}$-th block is the one before the final activation layer (softmax layer to generate multi-class probability maps). During the search process, $\mathcal{N}$ and spatial resolution can be determined within certain ranges of discrete integer values. The connection between previous blocks and the current one can be arbitrary. In order to further reduce GPU memory consumption during training and meanwhile speed-up the training process, we have one stem layer to down-sample input volume by half of its original size using 3 × 3 × 3 convolution following [43]. At the end of the architecture, another up-sampling layer (linear interpolation) is used to restore the feature maps back to the original volume size.

3.2. Encoding

To represent architecture and other training configurations and simplify the computation of the next step, we encode architecture and training configuration in the search space jointly to form a "large" one-dimensional vector $\mathcal{V}$. $\mathcal{V}$ encodes both numerical values and non-numerical values (*e.g.*, choices of optimizers/losses/data augmentation).

The architecture is encoded as a one-dimensional vector $\mathcal{A}$ with dynamic length. For each block, we use five integer indices to represent its current block ID, choice of operations, spatial resolution level, and two IDs of predecessor blocks, respectively. The IDs of predecessor for the 1st block is $(-1, -1)$, for the 2nd block is $(0, -1)$.

During training, we apply $n = 5$ augmentation methods in a sequence. Therefore, we have n place holders for m augmentation candidates. For each place holder, we use indices (0 to $m - 1$) to indicate the choice of augmentation method. Thus, the 1D vector for augmentation has length n. Meanwhile, we encode the options of different optimizers and loss functions using integer indices as well. The other related hyper-parameters (*e.g.*, learning rate) can be further optimized as long as they can be formulated into continuous or discrete values. After encoding all necessary components in the search space, we concatenate all 1D vectors into one large vector v.

Search space: The search space is designed to cover most components in a typical deep learning framework. For data augmentation, the candidates are random flipping (along $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ axes respectively), random rotation (90 degrees) in $\mathcal{X}$-$\mathcal{Y}$ planes, random zooming, random Gaussian noise, random intensity shift, random intensity scale shift. By default, we set the probability of activation (of each augmentation) to 0.15 according to the recommendations in [67]. The candidate learning rates are $[0.01, 0.005, 0.001, 0.0005, 0.0001]$, and the candidate learning rate schedulers include constant and polynomial

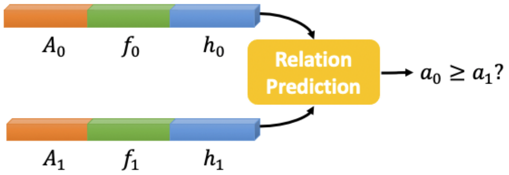

Figure 3. The predictor takes input vectors v_0 and v_1 (containing encodings of architecture A, augmentation f, and hyper-parameters h), and predicts which vector would generate higher validation scores a.

scheduler. The loss function candidates can be determined from (soft) Dice loss [19] with or without squared prediction, cross entropy (CE) loss, combinations of Dice loss and CE loss, and combinations of Dice loss and focal loss [68, 69]. The optimizer candidates are Adam, stochastic gradient descent, momentum, Nesterov and NovoGrad [70] optimizers. For the architecture space, the number of blocks $\mathcal{N}$ can be selected from 5 to 12 and the spatial resolution level l can be 2 to 5. At each spatial level, the spatial size of the feature maps is $1/2^{(l-1)}$ and the number of channels is $2^{(l-1)\cdot c_1}$. Here, c_1 is set to be 16. The block candidates are 3D residual block, 3D bottleneck block, and 3D axial-attention blocks.

3.3. Neural Predictor for Binary Relationship

The performance predictor takes the 1D encoding vector $\mathcal{V}$ containing neural architecture, data augmentation, and hyper-parameters, and outputs the corresponding performance (validation accuracy, etc.). Such a predictor is able to cover all possible components in machine/deep learning, and it can be potentially transferred across different datasets, tasks, and hardware platforms. However, the main drawback of the predictor-type AutoML methods is that they require a lot of training time to generate ground truth in order to train a stable predictor.

To alleviate the burden of training many jobs/instances, we proposed a new predictor-based search method to predict the relation between different configuration vectors v_i and v_j instead. The goal is to predict the relation of validation accuracy a between two configurations g_i and g_j: better or worse (*i.e.*, larger or smaller accuracy) shown in Fig. 3. After vector v is extracted from the raw configuration, we can adopt transformer modules [56] and fully-connected (FC) layers mapping the vector to a binary prediction. The transformer encoder is adopted to encode the entire vector v with dynamic length into feature maps with a fixed size. Multiple FC layers convert the high-level feature maps into binary relation predictions. The ground truth $\mathrm{GT}_{i,j}$ for training such a predictor is based on better or worse validation scores a_i, a_j as in Eq. 1.

$$\mathrm{GT}_{v_i,v_j} = \begin{cases} 1 & a_i \geq a_j \\ 0 & a_i < a_j \end{cases} \tag{1}$$

We formulate the predictor as a binary classifier instead of an accuracy regressor. Once the predictor is trained, it can be used to rank unseen configurations with a sorting algorithm. For each configuration, we compare it with all other sampled configurations. It would be indexed with how many configurations have worse validation accuracy than it. And sorting can be simply conducted based on such indices. The comparison between v_i and v_j is light-weighted using CPU or GPU. Thus, the sorting would be finished in seconds for hundreds of randomly sampled candidates.

The advantage of this predictor design is that it requires less overall training time. Firstly, various configurations can be compared with fewer training iterations, since the absolute values of predicted accuracy are not always necessary. Especially when transferring the predictor to another dataset, the predicted accuracy becomes much less informative. The actual ranking between configuration v_i and v_j is more informative to a new task without any search/training experience, compared to the predicted accuracy with the previous predictor. Second, achieving a stable predictor (accuracy regressor) requires at least hundreds of ground truth (GT) data points, which also increases the overall time cost [49]. In our method, the binary relation predictor requires much fewer GT points in order to learn a similar amount of parameters of the predictors. For instance, training 20 jobs can create only 20 GT points for an accuracy-based predictor. However, the same amount of trained jobs/instances can create $20 \times 20 = 400$ GT points for the relation-based predictor. Therefore, such a relation can be estimated with less training time/iterations.

4. Experiments

4.1. Datasets

LiTS 2017: The LiTS challenge is about developing automatic segmentation algorithms to segment liver lesions in contrast-enhanced abdominal CT scans [71]. It has been publicly available since 2017, and more than 3000 participants in the challenges made their submissions over the years. CT volumes and ground truth labels are provided by multiple clinical sites around the world. The dataset contains 201 3D abdominal CT volumes with ground truth labels of liver and lesion segmentation. 131 of them are used for model training, and 70 volumes are used for testing on a public server. The ground truth labels of the test set are not visible to the participants. The spacing of 3D CT volumes varies from $0.598 \times 0.598 \times 0.450$ mm^3 to $0.977 \times 0.977 \times 6.0$ mm^3, and the range of volume shapes is from $512 \times 512 \times 42$ voxel^3 to $512 \times 512 \times 1026$ voxel^3. Large variance also exists in the imaging quality, contrast, and field-of-view (FOV) (shown in Fig. 1). The dataset has been largely adopted for performance evaluation of 3D segmentation methods by many established works. In practice, the image intensity is clipped within an abdominal window of $[-125, 225]$ Hounsfield units (HU) and pre-processed using standard normalization. The dataset is re-sampled into the voxel spacing $(\min(1.0, s_x)\ mm, \min(1.0, s_y)\ mm, 1.0\ mm)$ based on each case's spacing s. Such re-sampling strategies are proved to be capable of preserving some small lesions in the high-resolution CT images.

Medical Segmentation Decathlon: The medical decathlon challenge (MSD) hosts several tasks of 3D medical image segmentation [72]. We choose to take 06 (lung lesion/tumor segmentation in 3D CT images) to test the transferability of our searched configurations from previous LiTS datasets. The dataset contains 63 volumes for training and validation, and 32 volumes for testing. The spacing of 3D CT volumes varies from $0.598 \times 0.598 \times 0.625\ mm^3$ to $0.977 \times 0.977 \times 2.5\ mm^3$, and the range of volume shapes is from $512 \times 512 \times 84$ to $512 \times 512 \times 947$. The image intensity is clipped in a chest window of $[-1000, 1000]$ HU and pre-processed using linear mapping to $[0, 1]$. The dataset is re-sampled into the isotropic voxel spacing 1.0 mm.

Implementation: The task for searching is the liver and lesion segmentation on LiTS dataset. The segmentation model takes a 1-channel input and outputs 3-class probability maps (background, liver, and lesion, respectively) with the same shape as the input. The task 06 of MSD is utilized to verify the transferability of our searched deep learning configurations from LiTS. The task is formulated as a binary segmentation problem. The last layer of search network is modified to take a 1-channel input and output 2-class probability maps (background, lung lesion).

To train and validate the predictor, we uniformly sample 100 configuration candidates from the search space. 75 of them are used for predictor training, and the remaining 25 are used for validation. Then, we train each candidate configuration for 10,000 iterations, and validate the segmentation model every 1,000 iterations. The best validation Dice score is referred to as the GT accuracy for that configuration. Once all the GT points are generated, the predictor model would be trained for 10,000 iterations with Adam optimizer, learning rate 0.001, and batch size 32. At last, we select the optimal configuration from the predictor with 200 candidates (100 existing ones and another 100 unseen random samples) to be our final model training solution. We use single GPU training jobs/instances for the configuration search. Each job takes about 3 hours for both training and validation, and the total searching time is around 300 GPU hours for the task. But our search process can be fully parallel, and thus searching would be done within two days using an 8-GPU server. Our searching efficiency is decent considering the huge search space. For example, C2FNAS used thousands of GPU hours to search for neural architectures (3D segmentation network) only [63]. The predictor model takes few minutes to train.

Our searched neural network is shown in Fig. 2. It has the least amount of parameters (16.96M), compared with other popular network models (19.07M for 3D U-Net [3], 22.60M for 3D U-Net++ [29], 17.02M for C2FNAS [63]). We reuse the configuration for Tab. 1 from the search method to determine all necessary components. During training, the input of the network are patches with size $128 \times 128 \times 128$, and the ratio between foreground and background patches is $1:1$. The batch size is 4 (2 patches from 2 volumes) per GPU. To achieve better and robust segmentation performance, we linearly extend the number of total training iterations to 40,000 with the same learning rate scheduler as the searched one. The validation is conducted per 1,000 iterations to select the best model checkpoint. The validation accuracy is measured with the Dice score. The model inference uses a sliding-window scheme, and the overlapped region of adjacent windows is 80% of the window size. Similar to [28, 67], we conducted 5-fold cross-validation for all tasks, resulting in 5 segmentation models after training. The final prediction of test data is the ensemble result of the probability maps from the 5 models. Our proposed method is implemented with PyTorch and trained on two NVIDIA V100 GPUs with 16 GB memory.

component	**searched result**
data augmentation	random flip, random gaussian noise, random intensity shift, random zoom, random intensity scale shift
learning rate (LR)	0.0001
LR scheduler	constant
loss function	Dice loss (no squared prediction) + CE
optimzer	Adam

Table 1. The final searched training configuration for liver and lesion segmentation using LiTS dataset.

4.2. Results

Evaluation: We evaluate the segmentation accuracy in terms of the Dice score per case for LiTS challenge, and the Dice score and the normalized surface distance (NSD) for the lung task of the MSD challenge. Tab. 2, Tab. 3 and Fig. 4 results demonstrate the superior performance of our method compared to the other SOTA methods.

For the LiTS dataset, our method performs better for both lesion and liver segmentation. Tab. 2 shows that our method was able to capture not only the lesion regions, but also the fine boundary of the large organs. Unlike other works, our method does not use any advanced model training features, such as sophisticated boundary loss functions [73] or deep supervision [28]. Our method benefits from fundamental components of neural network models and the training process. For the lung task of the MSD challenge, our method performs better than other SOTA methods with substantial improvements in terms of both the Dice score and the NSD. In the evaluation system of MSD, the higher the NSD value is, the better the segmentation quality is provided by the models. By inspecting the MSD leaderboard statistics of other methods for failed cases, we observe that all previous methods tend to make false negative predictions in a few specific test volumes (because the min metric values are close to 0). Whereas our method succeeds in such cases by capturing lesion regions well. Moreover, our method has clear advantage in 25 percentile and standard deviation of overall metric values, which shows our proposed method has both accurate and robust performance in this task. The fact that we migrated the found solution from LiTS to MSD lung task without any additional search, validates the transferability of our method.

	lesion	liver	mean
RA-UNet [74]	0.5950	0.9610	0.7780
AH-Net [75]	0.6340	0.9630	0.7895
FCN [76]	0.6610	0.9510	0.8060
3D-DenseUNet [77]	0.6960	0.9620	0.8290
H-DenseUNet [20]	0.7220	0.9610	0.8415
LW-HCN [78]	0.7300	0.9650	0.8475
U^3-Net [79]	0.7369	0.9638	0.8504
Cascade U-ResNets [80]	0.7520	0.9490	0.8505
VolumetricAttention [81]	0.7410	0.9610	0.8510
MA-Net [82]	0.7490	0.9600	0.8545
DistanceMetric [73]	0.7640	0.9650	0.8645
nnU-Net [67]	0.7630	0.9670	0.8650
T-AutoML (ours)	**0.7650**	**0.9670**	**0.8660**

Table 2. LiTS challenge test-set performance evaluation for lesion and liver segmentations in terms of the average Dice score per case. The metrics of our method are copied from the LiTS leaderboard, and the metrics of the other methods are copied from their respective publications and the leaderboard entries.

Comparison with nnU-Net: nnU-Net has been validated as the state-of-the-art method for several medical image segmentation tasks [28, 67]. It leverages 3 different types of 2D/3D U-Net architectures for prediction with ensembling. It requires training 15 models (3 networks for 5-fold cross validation) in parallel to determine which models are used to create an ensemble and generate the final prediction. Although we also conducted 5-fold cross validation using the searched architecture and training configuration, only one fixed neural architecture needs to be trained on each fold for a new task. This is because once the searching on a single task is accomplished, it can be applied to other similar tasks directly. Moreover, nnU-Net requires deep supervision during model training, which is unnecessary for our method. Therefore, the marginal effort of applying our searched model and training configuration to a new task is considerable less than nnU-Net. At the same time, our method achieved better performance and the overall segmentation quality.

	Dice			
	mean	std	min.	25%
MPUnet [83]	0.5900	-	-	-
C2FNAS [63]	0.7044	0.2099	0.0022	0.6042
nnU-Net [67]	0.7397	0.2164	0.0000	0.6041
T-AutoML (ours)	**0.7533**	**0.1574**	**0.3530**	**0.7014**
	NSD			
	mean	std	min.	25%
MPUnet [83]	0.5600	-	-	-
C2FNAS [63]	0.7222	0.2897	0.0031	0.5116
nnU-Net [67]	0.7602	0.2962	0.0000	0.7018
*U-Net++ [29]	0.7721	-	-	-
T-AutoML (ours)	**0.7768**	**0.2816**	**0.0998**	**0.7392**

Table 3. MSD challenge performance evaluation for lung tumor segmentation in terms of the Dice score and the normalized surface distance (NSD, higher is better). The metrics of our method are copied from the MSD leaderboard, and the metrics of other methods are copied from their respective publications and the leaderboard. *The results of U-Net++ are from the same task, but using a much larger training dataset.

4.3. Ablation Studies and Discussion

We conduct ablation studies to motivate our predictor choice (see Fig. 5). For fair comparisons, we utilize three different types of predictors to rank data points in the validation set. The accuracy-based predictor generates accuracy for each validation data, and uses the accuracy to rank them. The relation and transformer based relation predictor generates ranks after sorting. The regular relation predictor is based on a multi-layer perceptron (MLP). From Fig. 5, we see the accuracy predictor cannot produce a positive correlation (-0.322) between ground truth ranks and predicted ranks (based on linear regression). But both relation-based predictors can rank them in a positively correlated fashion. Especially the relation-based predictor captures the top ranking points (left bottom corner of Fig. 5). And the transformer-based predictor has more than 20% improvement over the MLP-based predictor in terms of the corre-

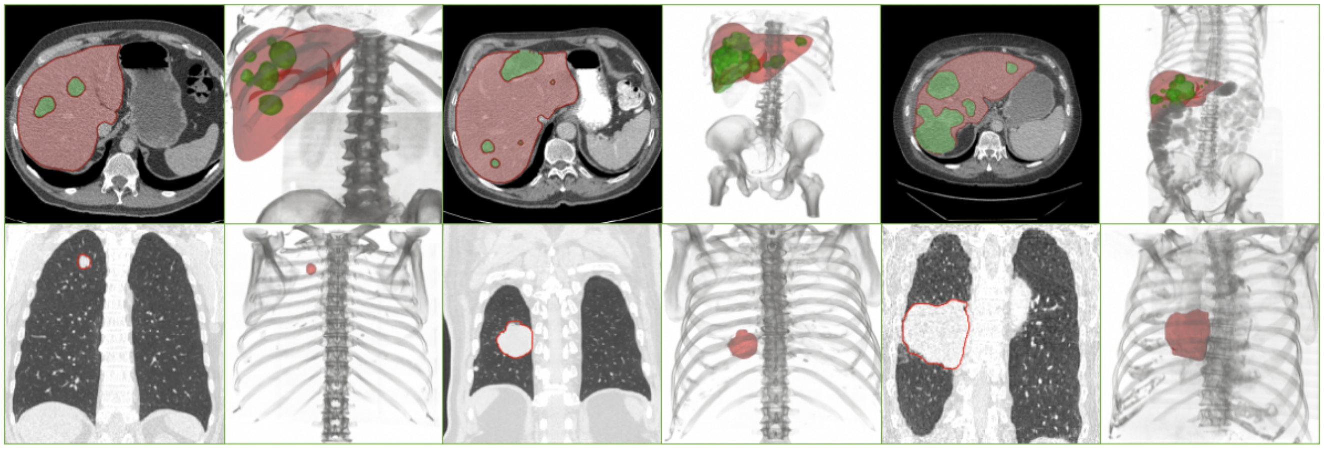

Figure 4. The first row shows the superimposed display of the liver and lesion segmentation in CT and their 3D rendering results. The second row shows the superimposed display of lung lesion segmentation in chest CT and its 3D rendering.

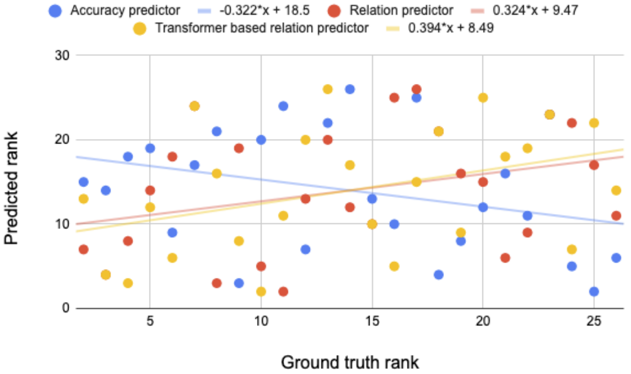

Figure 5. Comparison of different predictors for ranking prediction in validation set with 25 uniformly sampled configurations.

lation strength because of the advanced capacities of transformer modules. Because the accuracy predictor has much fewer GT points used for training, the training set is easily overfitted, resulting in poor validation performance. But the relation-based predictors leverage the pair-wise information (much more than individual accuracy) from the same training pool, and therefore are able to produce an improved ranking performance.

Based on the results of LiTS, apart from lesion segmentation, our method also works very well on organ segmentation. However, search on organ segmentation may not be able to generate good architecture or training configurations for lesion segmentation (even in the same dataset). This is because organ body segmentation is much simpler than lesion segmentation (especially for healthy cases). And organ segmentation models typically converge much faster than lesion segmentation models. Most of the segmentation networks with simple settings would achieve decent performance, even without data augmentation. The liver segmentation model would converge within 2,000 iterations via a constant learning rate 0.001, and its validation accuracy achieves a Dice score above 0.950. However, using the same dataset, the lesion segmentation model would take more iterations (around 10,000) to converge and data augmentation is necessary. Because of the relative simplicity of liver segmentation, searched configuration did not fully exploit the potentials of models and training recipes. Therefore, searching on liver segmentation would result in an AutoML predictor with low discriminatory capacity among candidates. Hence, we suspect that challenging tasks (such as lesion segmentation) are more helpful for AutoML search in order to transfer the searched results to other applications.

5. Conclusion

In this paper, we proposed a novel AutoML method to optimize deep learning configurations for lesion segmentation in 3D medical images leveraging the advanced capacity of transformer modules. The proposed method predicts the relation between different training configurations and neural networks. We introduced a new search space for the neural architecture through modifications of SpineNet [58]. Meanwhile, we created a predictor-based AutoML algorithm, which is computationally efficient and effective. It can cover "almost" all components in conventional deep learning frameworks. Our experiments showed that, compared to other existing methods in the literature, our method can achieve the most advanced performance in large-scale lesion segmentation datasets. Furthermore, the proposed methods have been proven to be effectively transferable to different datasets. Future work could investigate how to further reduce searching and training time given certain computational budgets.

References

[1] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 2021.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[3] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In Sébastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gözde B. Ünal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*, volume 9901 of *Lecture Notes in Computer Science*, pages 424–432, 2016.

[4] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[5] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8697–8710. IEEE Computer Society, 2018.

[6] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[8] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[9] Sumiaki Matsumoto, Harold L. Kundel, James C. Gee, Warren B. Gefter, and Hiroto Hatabu. Pulmonary nodule detection in ct images with quantized convergence index filter. *Medical Image Analysis*, 10(3):343–352, 2006.

[10] PC Vos, JO Barentsz, N Karssemeijer, and HJ Huisman. Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis. *Physics in Medicine & Biology*, 57(6):1527, 2012.

[11] K. Murphy, B. van Ginneken, A.M.R. Schilham, B.J. de Hoop, H.A. Gietema, and M. Prokop. A large-scale evaluation of automatic pulmonary nodule detection in chest ct using local image features and k-nearest-neighbour classification. *Medical Image Analysis*, 13(5):757–770, 2009.

[12] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Computer-aided detection of prostate cancer in mri. *IEEE transactions on medical imaging*, 33(5):1083–1092, 2014.

[13] Nathan Lay, Yohannes Tsehay, Yohan Sumathipala, Ruida Cheng, Sonia Gaur, Clayton Smith, Adrian Barbu, Le Lu, Baris Turkbey, Peter L Choyke, et al. A decomposable model for the detection of prostate cancer in multi-parametric mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 930–939. Springer, 2018.

[14] Jin Tae Kwak, Sheng Xu, Bradford J Wood, Baris Turkbey, Peter L Choyke, Peter A Pinto, Shijun Wang, and Ronald M Summers. Automated prostate cancer detection using t2-weighted and high-b-value diffusion-weighted magnetic resonance imaging. *Medical physics*, 42(5):2368–2378, 2015.

[15] Patrick Schelb, Simon Kohl, Jan Philipp Radtke, Manuel Wiesenfarth, Philipp Kickingereder, Sebastian Bickelhaupt, Tristan Anselm Kuder, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, et al. Classification of cancer at prostate mri: deep learning versus clinical pi-rads assessment. *Radiology*, 293(3):607–617, 2019.

[16] Tianjiao Liu, Qianqian Guo, Chunfeng Lian, Xuhua Ren, Shujun Liang, Jing Yu, Lijuan Niu, Weidong Sun, and Dinggang Shen. Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks. *Medical Image Analysis*, 58:101555, 2019.

[17] Marysia Winkels and Taco S. Cohen. Pulmonary nodule detection in ct scans with equivariant cnns. *Medical Image Analysis*, 55:15–26, 2019.

[18] Shuo Wang, Mu Zhou, Zaiyi Liu, Zhenyu Liu, Dongsheng Gu, Yali Zang, Di Dong, Olivier Gevaert, and Jie Tian. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical Image Anal.*, 40:172–183, 2017.

[19] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 565–571. IEEE Computer Society, 2016.

[20] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from CT volumes. *IEEE Trans. Medical Imaging*, 37(12):2663–2674, 2018.

[21] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017.

[22] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[23] Dong Yang, Daguang Xu, S Kevin Zhou, Bogdan Georgescu, Mingqing Chen, Sasa Grbic, Dimitris Metaxas, and Dorin Comaniciu. Automatic liver segmentation using an adversarial image-to-image network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 507–515. Springer, 2017.

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[25] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer, 2018.

[26] Siqi Liu, Daguang Xu, S Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 851–858. Springer, 2018.

[27] Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access*, 7:44247–44257, 2019.

[28] Fabian Isensee, Jens Petersen, André Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian J. Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *CoRR*, abs/1809.10486, 2018.

[29] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer F. Syeda-Mahmood, Anne L. Martel, Lena Maier-Hein, João Manuel R. S. Tavares, Andrew P. Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi, editors, *Deep Learning in Medical Image Analysis - and - Multimodal Learning for Clinical Decision Support - 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings*, volume 11045 of *Lecture Notes in Computer Science*, pages 3–11. Springer, 2018.

[30] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *arXiv preprint arXiv:1808.05377*, 2018.

[31] Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4092–4101. PMLR, 2018.

[32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[33] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 113–123. Computer Vision Foundation / IEEE, 2019.

[34] Dong Yang, Holger Roth, Ziyue Xu, Fausto Milletari, Ling Zhang, and Daguang Xu. Searching learning strategy with reinforcement learning for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–11. Springer, 2019.

[35] Irwan Bello, Barret Zoph, Vijay Vasudevan, and Quoc V Le. Neural optimizer search with reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 459–468. JMLR. org, 2017.

[36] Liang-Chieh Chen, Maxwell Collins, Yukun Zhu, George Papandreou, Barret Zoph, Florian Schroff, Hartwig Adam, and Jon Shlens. Searching for efficient multi-scale architectures for dense image prediction. In *Advances in Neural Information Processing Systems*, pages 8699–8710, 2018.

[37] Haowen Xu, Hao Zhang, Zhiting Hu, Xiaodan Liang, Ruslan Salakhutdinov, and Eric Xing. Autoloss: Learning discrete schedules for alternate optimization. *arXiv preprint arXiv:1810.02442*, 2018.

[38] Nanqing Dong, Min Xu, Xiaodan Liang, Yiliang Jiang, Wei Dai, and Eric Xing. Neural architecture search for adversarial medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 828–836. Springer, 2019.

[39] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7036–7045, 2019.

[40] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018.

[41] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architecture search. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[42] Yingda Xia, Fengze Liu, Dong Yang, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger

Roth. 3d semi-supervised learning with uncertainty-aware multi-view co-training. *arXiv preprint arXiv:1811.12506*, 2018.

[43] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. *arXiv preprint arXiv:1901.02985*, 2019.

[44] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. *arXiv preprint arXiv:1904.00420*, 2019.

[45] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10734–10742, 2019.

[46] Albert Shaw, Daniel Hunter, Forrest Landola, and Sammy Sidhu. Squeezenas: Fast neural architecture search for faster semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[47] Saining Xie, Alexander Kirillov, Ross Girshick, and Kaiming He. Exploring randomly wired neural networks for image recognition. *arXiv preprint arXiv:1904.01569*, 2019.

[48] Xinyu Gong, Shiyu Chang, Yifan Jiang, and Zhangyang Wang. Autogan: Neural architecture search for generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3224–3234, 2019.

[49] Tianzhe Wang, Kuan Wang, Han Cai, Ji Lin, Zhijian Liu, Hanrui Wang, Yujun Lin, and Song Han. APQ: joint search for network architecture, pruning and quantization policy. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2075–2084. IEEE, 2020.

[50] Lukasz Dudziak, Thomas C. P. Chau, Mohamed S. Abdelfattah, Royson Lee, Hyeji Kim, and Nicholas D. Lane. BRP-NAS: prediction-based NAS using gcns. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[51] Jiahui Yu, Pengchong Jin, Hanxiao Liu, Gabriel Bender, Pieter-Jan Kindermans, Mingxing Tan, Thomas S. Huang, Xiaodan Song, Ruoming Pang, and Quoc Le. Bignas: Scaling up neural architecture search with big single-stage models. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII*, volume 12352 of *Lecture Notes in Computer Science*, pages 702–717. Springer, 2020.

[52] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 4780–4789, 2019.

[53] Martin Wistuba and Tejaswini Pedapati. Learning to rank learning curves. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10303–10312. PMLR, 2020.

[54] Gaurav Mittal, Chang Liu, Nikolaos Karianakis, Victor Fragoso, Mei Chen, and Yun Fu. Hyperstar: Task-aware hyperparameters for deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8733–8742. IEEE, 2020.

[55] Haowen Xu, Hao Zhang, Zhiting Hu, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Autoloss: Learning discrete schedules for alternate optimization. *CoRR*, abs/1810.02442, 2018.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[57] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *Deep learning and data labeling for medical applications*, pages 179–187. Springer, 2016.

[58] Xianzhi Du, Tsung-Yi Lin, Pengchong Jin, Golnaz Ghiasi, Mingxing Tan, Yin Cui, Quoc V. Le, and Xiaodan Song. Spinenet: Learning scale-permuted backbone for recognition and localization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11589–11598. IEEE, 2020.

[59] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L. Yuille, and Fei-Fei Li. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 82–92. Computer Vision Foundation / IEEE, 2019.

[60] Woong Bae, Seungho Lee, Yeha Lee, Beomhee Park, Minki Chung, and Kyu-Hwan Jung. Resource optimized neural architecture search for 3d medical image segmentation. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali R. Khan, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part II*, volume 11765 of *Lecture Notes in Computer Science*, pages 228–236. Springer, 2019.

[61] Sungwoong Kim, Ildoo Kim, Sungbin Lim, Woonhyuk Baek, Chiheon Kim, Hyungjoo Cho, Boogeon Yoon, and Taesup Kim. Scalable neural architecture search for 3d medical image segmentation. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali R. Khan, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part III*, volume 11766 of *Lecture Notes in Computer Science*, pages 220–228. Springer, 2019.

[62] Zhuotun Zhu, Chenxi Liu, Dong Yang, Alan L. Yuille, and Daguang Xu. V-NAS: neural architecture search for volumetric medical image segmentation. In *2019 International Conference on 3D Vision, 3DV 2019, Québec City, QC, Canada, September 16-19, 2019*, pages 240–248. IEEE, 2019.

[63] Qihang Yu, Dong Yang, Holger Roth, Yutong Bai, Yixiao Zhang, Alan L. Yuille, and Daguang Xu. C2FNAS: coarse-to-fine neural architecture search for 3d medical image segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4125–4134. IEEE, 2020.

[64] Yuanfeng Ji, Ruimao Zhang, Zhen Li, Jiamin Ren, Shaoting Zhang, and Ping Luo. Uxnet: Searching multi-level feature aggregation for 3d medical image segmentation. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I*, volume 12261 of *Lecture Notes in Computer Science*, pages 346–356. Springer, 2020.

[65] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan L. Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 108–126. Springer, 2020.

[66] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[67] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[68] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society, 2017.

[69] Wentao Zhu, Yufang Huang, Hui Tang, Zhen Qian, Nan Du, Wei Fan, and Xiaohui Xie. Anatomynet: Deep 3d squeeze-and-excitation u-nets for fast and fully automated whole-volume anatomical segmentation. *bioRxiv*, page 392969, 2018.

[70] Boris Ginsburg, Patrice Castonguay, Oleksii Hrinchuk, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, Huyen Nguyen, and Jonathan M. Cohen. Stochastic gradient methods with layer-wise adaptive moments for training of deep networks. *CoRR*, abs/1905.11286, 2019.

[71] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, Samuel Kadoury, Tomasz K. Konopczynski, Miao Le, Chunming Li, Xiaomeng Li, Jana Lipková, John S. Lowengrub, Hans Meine, Jan Hendrik Moltz, Chris Pal, Marie Piraud, Xiaojuan Qi, Jin Qi, Markus Rempfler, Karsten Roth, Andrea Schenk, Anjany Sekuboyina, Ping Zhou, Christian Hülsemeyer, Marcel Beetz, Florian Ettlinger, Felix Grün, Georgios Kaissis, Fabian Lohöfer, Rickmer Braren, Julian Holch, Felix Hofmann, Wieland H. Sommer, Volker Heinemann, Colin Jacobs, Gabriel Efrain Humpire Mamani, Bram van Ginneken, Gabriel Chartrand, An Tang, Michal Drozdzal, Avi Ben-Cohen, Eyal Klang, Michal Marianne Amitai, Eli Konen, Hayit Greenspan, Johan Moreau, Alexandre Hostettler, Luc Soler, Refael Vivanti, Adi Szeskin, Naama Lev-Cohain, Jacob Sosna, Leo Joskowicz, and Bjoern H. Menze. The liver tumor segmentation benchmark (lits). *CoRR*, abs/1901.04056, 2019.

[72] Medical decathlon challenge, 2018.

[73] Yi Zhang, Xiwen Pan, Congsheng Li, and Tongning Wu. 3d liver and tumor segmentation with cnns based on region and distance metrics. *Applied Sciences*, 10(11):3794, 2020.

[74] Qiangguo Jin, Zhao-Peng Meng, Changming Sun, Leyi Wei, and Ran Su. Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in CT scans. *CoRR*, abs/1811.01328, 2018.

[75] Siqi Liu, Daguang Xu, Shaohua Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna K. Jerebko, Weidong Cai, and Dorin Comaniciu. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, volume 11071 of *Lecture Notes in Computer Science*, pages 851–858. Springer, 2018.

[76] Eugene Vorontsov, An Tang, Chris Pal, and Samuel Kadoury. Liver lesion segmentation informed by joint liver segmentation. In *15th IEEE International Symposium on Biomedical Imaging, ISBI 2018, Washington, DC, USA, April 4-7, 2018*, pages 1332–1335. IEEE, 2018.

[77] Nasser Alalwan, Amr Abozeid, AbdAllah A ElHabshy, and Ahmed Alzahrani. Efficient 3d deep learning model for med-

ical image semantic segmentation. *Alexandria Engineering Journal*, 60(1):1231–1239, 2021.

[78] Jianpeng Zhang, Yutong Xie, Pingping Zhang, Hao Chen, Yong Xia, and Chunhua Shen. Light-weight hybrid convolutional network for liver tumor segmentation. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4271–4277. ijcai.org, 2019.

[79] Song-Toan Tran, Ching-Hwa Cheng, and Don-Gey Liu. A multiple layer u-net, u^{n}-net, for liver and liver tumor segmentation in CT. *IEEE Access*, 9:3752–3764, 2021.

[80] Xue-Feng Xi, Lei Wang, Victor S. Sheng, Zhiming Cui, Baochuan Fu, and Fuyuan Hu. Cascade u-resnets for simultaneous liver and lesion segmentation. *IEEE Access*, 8:68944–68952, 2020.

[81] Xudong Wang, Shizhong Han, Yunqiang Chen, Dashan Gao, and Nuno Vasconcelos. Volumetric attention for 3d medical image segmentation and detection. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali R. Khan, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part VI*, volume 11769 of *Lecture Notes in Computer Science*, pages 175–184. Springer, 2019.

[82] Tongle Fan, Guanglei Wang, Yan Li, and Hongrui Wang. Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access*, 8:179656–179665, 2020.

[83] Mathias Perslev, Erik Bjørnager Dam, Akshay Pai, and Christian Igel. One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali R. Khan, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part II*, volume 11765 of *Lecture Notes in Computer Science*, pages 30–38. Springer, 2019.