

# LAMP: Large Deep Nets with Automated Model Parallelism for Image Segmentation

Wentao Zhu, Can Zhao, Wenqi Li, Holger Roth, Ziyue Xu, and Daguang Xu

NVIDIA

**Abstract.** Deep Learning (DL) models are becoming larger, because the increase in model size might offer significant accuracy gain. To enable the training of large deep networks, data parallelism and model parallelism are two well-known approaches for parallel training. However, data parallelism does not help reduce memory footprint per device. In this work, we introduce Large deep 3D ConvNets with Automated Model Parallelism (LAMP) and investigate the impact of both input’s and deep 3D ConvNets’ size on segmentation accuracy. Through automated model parallelism, it is feasible to train large deep 3D ConvNets with a large input patch, even the whole image. Extensive experiments demonstrate that, facilitated by the automated model parallelism, the segmentation accuracy can be improved through increasing model size and input context size, and large input yields significant inference speedup compared with sliding window of small patches in the inference. Code is available<sup>1</sup>.

**Keywords:** Automated model parallelism · Large deep ConvNets · Large image segmentation · Parallel U-Net.

## 1 Introduction

Currently, deep learning models have been becoming larger. More and more studies demonstrate that, the increase in model size offers significant accuracy gain. In the natural language processing (NLP), transformers have paved the way for large models. For instance, the Bert-large model [7] consumes 0.3 billion (B) parameters and GPT-2 [18] has 1.5B parameters. In the image classification of computer vision, AmoebaNet (B) [10] consists of 550 million (M) parameters and achieves the best top-1 accuracy of 84.4% on ImageNet 2012 validation dataset [6]. As the model size continues to grow, training these large models becomes challenging because it is difficult to fit the training within the memory limit of one single GPU.

There are several ways to train large models on GPUs. Model compression, such as mixed precision training [16], tries to use less bits to represent the network. It can reduce GPU memory consumption to some extent, however, might affect accuracy and can only fit a slightly or moderately large model to one GPU. Checkpointing [4, 15] reduces the memory of the intermediate feature maps and

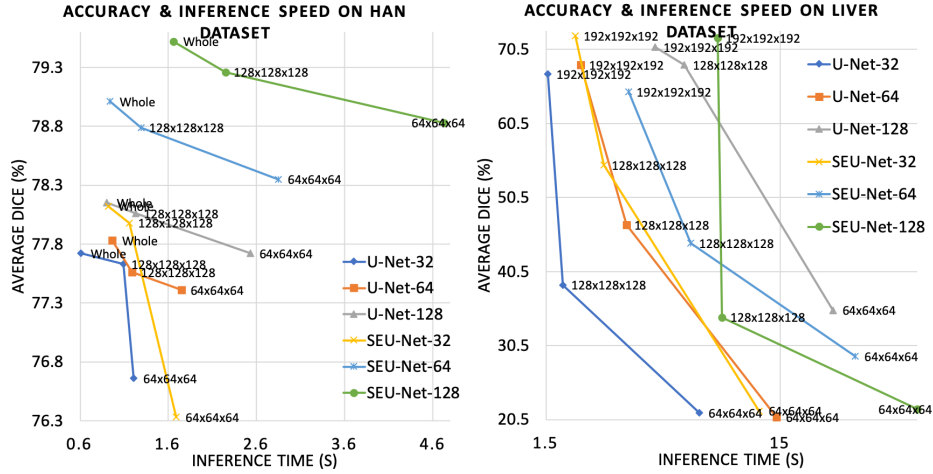
---

<sup>1</sup> <https://monai.io/research/lamp-automated-model-parallelism>

gradients during training, such that the memory consumption can be reduced to  $O(\log n)$  with  $O(n \log n)$  extra time for forward computation in the network of  $n$  layers theoretically. Invertible networks [8, 2, 3, 32] further reduce memory consumption to  $O(1)$  by modifying the networks to be invertible which recalculate the feature maps in the back-propagation and might impact accuracy for discriminative models such as commonly used U-Net for segmentation [21].

Facilitated by the high speed communication tools such as NVLINK, parallel training across devices is a popular direction for this challenge. Generally, there are two common parallelisms to fit large models into GPUs without information loss and re-calculation, data parallelism and model parallelism [10, 19, 17]. Data parallelism duplicates the model and runs split batch in multiple devices. It does not reduce model’s memory footprint per device and cannot address out of memory issue faced by training large models. Model parallelism splits a model into multiple partitions and naturally handles this issue. For instance, a state-of-the-art model parallelism, Megatron, can scale up to 20B parameter models by using 16 GPUs. Advanced model parallelism executes partitions concurrently across devices for efficient training, and multiple model parallelisms have emerged, e.g., pipeline parallelism in GPipe [10] and PipeDream [17], and TensorSlicing [22] in Megatron [23] and Mesh Tensorflow [22]. However, model parallelisms, such as Megatron [23], only support a limited set of operators and models. For example, in medical image analysis, the most widely used model, U-Net [21], is not supported by these existing parallelisms. In medical domain, it is a common need to be able to handle 3D volumetric image, which essentially consumes more memory with 3D ConvNets than their 2D counterparts. Unfortunately, current medical image computing is still limited by GPU memory size. A lot of techniques, such as sliding window and resampling, are utilized to get around the problem. Moreover, the designed 3D models often use much less filters than advanced 2D models in each convolution [11]. Therefore, insightful investigations of large models and large context, i.e., large input, might be extremely useful for the current research by leveraging automated model parallelism.

Training large models with large input is especially challenging for medical images due to limited number of training data. Large input increases context which is critical for image understanding [11]. However, it reduces the variation of training input and aggravates the extremely imbalance issue among background and relatively small subjects (e.g., small organs and lesions) commonly existed in medical image computing [29, 25]. Various loss functions have been proposed to alleviate this challenge. For example, adaptive weighted loss is proposed with a hybrid loss between dice loss of class-level loss and focal loss of voxel-level loss for small organ segmentation [29]. The second example is the boundary loss [13], which is different from previous approaches using unbalanced integrals over the regions. It uses integrals over the boundary (interface) between the regions, which can be implemented by a level set distance map weighted cross entropy loss leveraging an integral approach to computing boundary variations. Transfer learning by fine-tuning from a pretrained model is another way to reduce the training difficulty of specially designed medical image models [26]. Based on



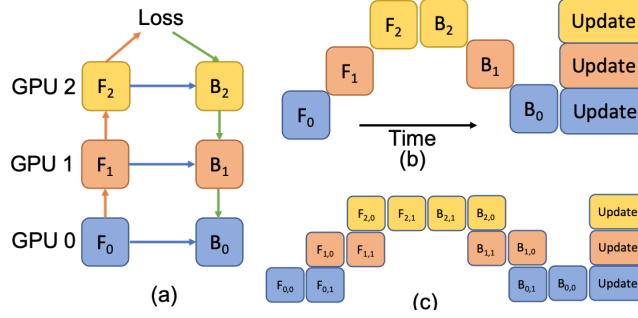
**Fig. 1.** Segmentation accuracy (Dice coefficient, %) and inference time (s) comparisons among 3D U-Net and 3D SEU-Net of different sizes (#filters in the first convolutional layer: 32, 64, 128) and different input sizes ( $64 \times 64 \times 64$ ,  $128 \times 128 \times 128$ , whole image or  $192 \times 192 \times 192$ ) on HaN nine organ auto-segmentation and decathlon liver and tumor segmentation datasets. Large model and input yield better segmentation accuracy consistently, and large input significantly decreases inference time.

learning theory such as curriculum learning [1, 12], a model can be well trained by firstly being fit easy samples/tasks and later being fit hard samples/tasks.

**Contributions** In this work, we investigate the impact of model size and input size in medical image analysis. We choose 3D U-Net [21] and the other advanced U-Net, 3D Squeeze-and-Excitation U-Net (SEU-Net) [9] in AnatomyNet [29], and validate them on large image segmentation tasks, i.e., head and neck (HaN) multi-organ segmentation [29] and decathlon liver and tumor segmentation [24]. Considering the flexibility and efficiency, we design a parallel U-Net based on GPipe [10] as the back-end parallelism. In the training, we employ existing well-designed adaptive weighted loss in [29] and design a curriculum training strategy based on different input sizes. Specifically, we sequentially fit the model with small patches for training in the first stage, medium patches thereafter, and large input lastly. We conduct extensive experiments, and conclude that, employing large models and input context increases segmentation accuracy. Large input also reduces inference time significantly by leveraging automated model parallelism in Fig. 1.

## 2 Method

Considering flexibility and efficiency, we employ GPipe [10] as the backend parallelism. The model parallelism is introduced in Section 2.1. We describe how to



**Fig. 2.** (a) A deep model is partitioned across three GPUs.  $F_k$  is the forward function of the  $k$ -th cell.  $B_k$  is the back-propagation function which relies on both  $B_{k+1}$  from upper layer and feature  $F_k$ . (b) Conventional model parallelism has low device utilization because of dependency of the model. (c) Pipeline parallelism splits the input mini-batch to smaller micro-batches (two micro-batches in the figure) and enables different devices to run micro-batches simultaneously. Synchronized gradient calculation can be applied lastly.

design a parallel U-Net in Section 2.2. How to train the large models with large context input is introduced in Section 2.3.

## 2.1 Automated Model Parallelism

Deep networks can be defined as a sequential model of  $L$  layers. Each layer  $L_i$  can be modeled by a forward computation function  $f_i$  with parameters  $w_i$ . Given the number of partitions  $K$ , i.e., the number of GPUs typically, the model can be partitioned into  $K$  parts as illustrated in Fig. 2 (a). Specifically, let part  $p_k$  consist of consecutive layers from layer  $L_i$  to layer  $L_j$ . The parameters of part  $p_k$  is the union of parameters  $w_i, w_{i+1}, \dots, w_j$ , and the forward function can be derived sequentially

$$F_k = f_j \circ f_{j-1} \circ \dots \circ f_i. \quad (1)$$

According to the chain rule in the gradient calculation, the back-propagation function  $B_k$  can be derived from  $F_k$  by automated symbolic differentiation in the existing deep learning packages, e.g., PyTorch.

In the forward pass, GPipe [10, 14] first splits the input mini-batch of size  $N$  to  $M$  micro-batches as illustrated in Fig 2 (c). Micro-batches are pipelined through  $K$  devices by model parallelism sequentially as illustrated in Fig 2 (b). This micro-batch splitting in Fig 2 (c) has a higher device utilization than conventional model parallelism in Fig 2 (b). After forward pass of all the micro-batches in the current mini-batch, gradients from all  $M$  micro-batches are accumulated synchronously and back-propagation is applied to update model parameters. GPipe reduces space complexity from  $O(N \times L)$  to  $O(N + \frac{L}{K} \times \frac{N}{M})$ , where  $\frac{L}{K}$  is the size of layers per partition and  $\frac{N}{M}$  is the micro-batch size [10].

## 2.2 Parallel U-Net

The pipeline parallelism is extremely simple and intuitive, and it is flexible and can be easily used to design various parallel algorithms. To use GPipe, we only need to 1) set the number of partitions  $K$ , which is the number of GPUs typically, 2) set the number of micro-batches  $M$ , which can also be set as the number of GPUs for efficiency, 3) modify the network into sequential layers. Next, we describe how to design a parallel U-Net.

We employ the conventional U-Net [21], which can be divided into three parts: an encoder  $E$  with five blocks  $e_1, e_2, \dots, e_5$  from input sequentially, a decoder  $D$  with four blocks  $d_5, d_4, \dots, d_1$ , and four skip connections  $s_1, s_2, \dots, s_4$ . The U-Net can be formulated

$$\begin{aligned} E &= e_5 \circ e_4 \circ \dots \circ e_1, & d_i &= s_i(e_i, d_{i+1}), i = 1, 2, \dots, 4, \\ D &= d_1 \circ d_2 \circ \dots \circ d_5, \end{aligned} \quad (2)$$

where  $s_i$  is typically a concatenation along channel dimension. The input of encoder  $E$  is the image, and the input of decoder block  $d_5$  is the output of encoder. We can then add a softmax function after decoder  $D$  for segmentation.

The main challenge of pipeline-based parallel U-Net is the dependency of intermediate encoder in the skip connection  $s_i$ . GPipe requires that the model needs to be implemented in a sequential way. However, each  $e_i, i = 1, 2, \dots, 4$ , is used in both encoder and decoder, which affects automated partition in GPipe. We can remove the dependency and modify U-Net by duplicating the output of each encoder  $e_i = \{e_{i,0}, e_{i,1}\}, i = 1, 2, \dots, 4$ . Specifically, the sequential U-Net can be derived

$$\begin{aligned} E &= e_5 \circ e_{4,0} \circ \dots \circ e_{2,0} \circ e_{1,0}, & d_i &= s_i(e_{i,1}, d_{i+1}), i = 1, 2, \dots, 4, \\ D &= d_1 \circ d_2 \circ \dots \circ d_5. \end{aligned} \quad (3)$$

The temporary variable  $e_{i,1}$  breaks the dependency in the skip connection and facilitates the automated partition in automated parallelism of GPipe. We can employ the existing GPipe algorithm to implement parallel U-Net based on the designed sequential U-Net.

## 2.3 Learning Large Models

Leveraging the powerful tool of parallel U-Net, we investigate the impact of model size and input context size. Although previous study demonstrates large input size increases segmentation accuracy because of large context [11], it also decreases the variation of training input and aggravates the extremely imbalance issue between background and the small subjects. From model size's perspective, large model consists of more parameters which typically require more various data to fit. Therefore, designing a learning strategy is essential to fully exploit the power of large input with more context information.

Inspired by the learning theory, i.e. curriculum learning [1], we can fit easy data/task into the network first and let the network to solve hard task later.

**Table 1.** Dice coefficient (%) achieved on the HaN test set using different sizes of U-Nets and inputs.

Models	BS	CH	MA	OL	OR	PL	PR	SL	SR	Average $\uparrow$
U-Net-32 ( $64^3$ )	84.23	48.87	89.75	69.11	68.28	87.43	85.48	79.36	77.41	76.66
U-Net-64 ( $64^3$ )	84.28	46.21	91.55	70.34	69.92	87.76	85.98	81.46	79.23	77.41
U-Net-128 ( $64^3$ )	84.58	48.52	91.12	71.04	69.28	87.76	85.78	81.34	80.03	77.72
U-Net-32 ( $128^3$ )	84.23	53.30	91.97	70.29	68.40	87.43	85.48	79.36	78.17	77.63
U-Net-64 ( $128^3$ )	84.71	46.21	92.47	70.34	69.92	87.76	85.98	81.46	79.23	77.56
U-Net-128 ( $128^3$ )	84.84	48.52	93.71	71.04	69.28	87.76	85.78	81.57	80.03	78.06
U-Net-32 (Whole)	84.23	53.30	91.97	70.29	68.40	87.43	85.48	79.36	79.02	77.72
U-Net-64 (Whole)	84.71	48.59	92.47	70.34	69.92	87.76	85.98	81.46	79.23	77.83
U-Net-128 (Whole)	84.84	48.52	93.71	71.04	70.09	87.76	85.78	81.57	80.03	78.15

**Table 2.** Dice coefficient (%) achieved on the HaN test set using different sizes of SEU-Nets and inputs.

Models	BS	CH	MA	OL	OR	PL	PR	SL	SR	Average $\uparrow$
AnatomyNet [29]	86.65	53.22	92.51	72.10	70.64	88.07	87.35	81.37	81.30	79.25
SEU-Net-32 ( $64^3$ )	84.07	47.09	90.12	68.58	69.73	87.14	85.21	79.20	75.81	76.33
SEU-Net-64 ( $64^3$ )	85.49	50.32	92.45	71.93	69.94	88.24	86.27	81.15	79.37	78.35
SEU-Net-128 ( $64^3$ )	86.38	51.85	93.55	70.62	70.08	88.11	85.99	81.79	81.13	78.83
SEU-Net-32 ( $128^3$ )	85.76	50.52	92.91	70.76	69.73	87.31	85.86	81.03	77.95	77.98
SEU-Net-64 ( $128^3$ )	85.73	50.37	94.26	71.97	71.09	88.34	86.58	81.15	79.64	78.79
SEU-Net-128 ( $128^3$ )	86.38	51.85	93.87	71.63	70.44	88.11	86.75	81.79	82.48	79.26
SEU-Net-32 (Whole)	85.76	51.27	92.91	70.76	69.73	87.31	85.86	81.03	78.43	78.12
SEU-Net-64 (Whole)	85.73	52.29	94.26	71.97	71.09	88.34	86.58	81.15	79.64	79.01
SEU-Net-128 (Whole)	86.38	51.85	93.87	73.70	70.44	88.26	86.75	81.96	82.48	<b>79.52</b>

Learning from smaller patches is easier, because smaller patches can be sampled with less imbalance and the lower dimension of smaller patches consists of less structures to learn for structured tasks, e.g., image segmentation. In practice, we firstly sample small positive patches (size of  $64 \times 64 \times 64$ ) to train the model in the initial stage. In the second stage, we sample medium positive patches (size of  $128 \times 128 \times 128$ ) to train the model. Finally, we use the largest patch to train the model. In this way, we can fully train models with large input patches in a practical way.

### 3 Experiments

We use two datasets to investigate the impact of large models and large input context for segmentation, the head and neck (HaN) and decathlon liver datasets. The HaN dataset consists of whole-volume computed tomography (CT) images with manually generated binary masks of nine anatomies, i.e., brain stem (BS), chiasm (CH), mandible (MD), optic nerve left (OL), optic nerve right (OR), parotid gland left (PL), parotid gland right (PR), submandibular gland left (SL),

**Table 3.** Average inference time (s) per test image achieved on the HaN test set using different sizes of networks and inputs.

Models	Inference time ↓	Models	Inference time ↓
U-Net-32 ( $64^3$ ) $2 \times 16G$	$1.21 \pm 0.07$	SEU-Net-32 ( $64^3$ ) $2 \times 16G$	$1.69 \pm 0.17$
U-Net-64 ( $64^3$ ) $4 \times 16G$	$1.75 \pm 0.08$	SEU-Net-64 ( $64^3$ ) $2 \times 32G$	$2.85 \pm 0.13$
U-Net-128 ( $64^3$ ) $2 \times 32G$	$2.53 \pm 0.04$	SEU-Net-128 ( $64^3$ ) $4 \times 32G$	$4.73 \pm 0.69$
U-Net-32 ( $128^3$ )	$1.09 \pm 0.28$	SEU-Net-32 ( $128^3$ )	$1.16 \pm 0.36$
U-Net-64 ( $128^3$ )	$1.19 \pm 0.16$	SEU-Net-64 ( $128^3$ )	$1.29 \pm 0.18$
U-Net-128 ( $128^3$ )	$1.23 \pm 0.16$	SEU-Net-128 ( $128^3$ )	$2.25 \pm 0.13$
U-Net-32 (Whole)	$0.61 \pm 0.07$	SEU-Net-32 (Whole)	$0.92 \pm 0.07$
U-Net-64 (Whole)	$0.96 \pm 0.22$	SEU-Net-64 (Whole)	$0.94 \pm 0.07$
U-Net-128 (Whole)	$0.90 \pm 0.14$	SEU-Net-128 (Whole)	$1.66 \pm 0.14$

and submandibular gland right (SR). We download the publicly available pre-processed data from AnatomyNet [29], which includes three public datasets: 1) MICCAI Head and Neck Auto Segmentation Challenge 2015 [20]; 2) the Head-Neck Cetuximab collection from The Cancer Imaging Archive (TCIA) [5]; 3) the CT images from four different institutions in Québec, Canada [28], also from TCIA. We use the dataset directly for fair comparison with benchmark methods. The dataset consists of 261 training images with missing annotations and ten test samples consisting of all annotations of nine organs. The largest image size can be  $352 \times 256 \times 288$ . We use the same data augmentation techniques in [29].

The other dataset is 3D liver and tumor segmentation CT dataset from the medical segmentation decathlon [24]. We randomly split the dataset into 104 training images and 27 test images. We re-sample the CT images to  $1 \times 1 \times 1$  mm<sup>3</sup> spacing. To focus on the liver region, we clip the voxel value within range  $[-21, 89]$  and linearly transform each 3D image into range  $[0, 1]$ . In the training, we randomly flip and rotation 90 degrees in XY space with probability 0.1. We further add uniform random noise  $[-0.2, 0.2]$  to augment the training data. The largest image size can be  $512 \times 512 \times 704$ . We will release the script and data splitting for reproducibility.

In the training, for the largest input, we use batch size of one and RMSProp optimizer [27] with 300 epochs and learning rate of  $1 \times 10^{-3}$ . For training with patch size  $128 \times 128 \times 128$ , we use batch size of four and 1200 epochs. For training with patch size  $64 \times 64 \times 64$ , we use batch size of 16 and 4800 epochs. For U-Net-32 and Squeeze-and-Excitation U-Net (SEU-Net-32), the number of filters in each convolution of the first encoder block is 32. We increase the number of filters to 64 and 128 to investigate the impact of increasing model size. In the encoder of each model, the number of filters are doubled with the increase of encoder blocks accordingly. The decoder is symmetric with the encoder.

We employ two networks, 3D U-Net and 3D SEU-Net, to investigate the impact of model size and input context size in table 1 and 2 on HaN dataset. With the increase of model size and input size, the segmentation accuracy increases consistently for both U-Net and SEU-Net. The SEU-Net-128 with whole image as input achieves better performance than AnatomyNet searching differ-

**Table 4.** Dice coefficientt (%) achieved on the Decathlon liver segmentation test set using different sizes of inputs and U-Nets and SEU-Nets.

Models	Liver	Tumor	Average $\uparrow$	Models	Liver	Tumor	Aevage $\uparrow$
U-Net-32 ( $64^3$ )	4.76	38.06	21.41	SEU-Net-32 ( $64^3$ )	0.73	42.56	21.65
U-Net-64 ( $64^3$ )	9.70	31.96	20.83	SEU-Net-64 ( $64^3$ )	11.90	46.19	29.05
U-Net-128 ( $64^3$ )	34.52	35.99	35.26	SEU-Net-128 ( $64^3$ )	0.34	43.44	21.89
U-Net-32 ( $128^3$ )	26.23	51.12	38.68	SEU-Net-32 ( $128^3$ )	58.88	50.83	54.86
U-Net-64 ( $128^3$ )	40.95	52.63	46.79	SEU-Net-64 ( $128^3$ )	38.38	50.25	44.32
U-Net-128 ( $128^3$ )	84.83	51.98	68.41	SEU-Net-128 ( $128^3$ )	20.20	48.44	34.32
U-Net-32 ( $192^3$ )	82.83	51.57	67.20	SEU-Net-32 ( $192^3$ )	89.25	55.38	72.32
U-Net-64 ( $192^3$ )	91.58	45.29	68.44	SEU-Net-64 ( $192^3$ )	77.66	51.93	64.80
U-Net-128 ( $192^3$ )	90.99	50.67	70.83	SEU-Net-128 ( $192^3$ )	87.61	56.48	72.05

**Table 5.** Average inference time (s) per test image achieved on the Decathlon liver segmentation test set using different sizes of networks and inputs.

Models	Inference time $\downarrow$	Models	Inference time $\downarrow$
U-Net-32 ( $64^3$ ) $2 \times 16G$	$6.78 \pm 0.06$	SEU-Net-32 ( $64^3$ ) $4 \times 16G$	$12.23 \pm 0.08$
U-Net-64 ( $64^3$ ) $4 \times 16G$	$14.52 \pm 0.02$	SEU-Net-64 ( $64^3$ ) $2 \times 32G$	$31.47 \pm 0.16$
U-Net-128 ( $64^3$ ) $4 \times 32G$	$25.37 \pm 1.10$	SEU-Net-128 ( $64^3$ ) $8 \times 32G$	$57.99 \pm 11.08$
U-Net-32 ( $128^3$ )	$1.77 \pm 0.42$	SEU-Net-32 ( $128^3$ )	$2.64 \pm 0.06$
U-Net-64 ( $128^3$ )	$3.30 \pm 0.52$	SEU-Net-64 ( $128^3$ )	$6.23 \pm 0.17$
U-Net-128 ( $128^3$ )	$5.84 \pm 0.21$	SEU-Net-128 ( $128^3$ )	$8.49 \pm 0.08$
U-Net-32 ( $256^3$ )	$1.52 \pm 0.58$	SEU-Net-32 ( $256^3$ )	$2.00 \pm 0.20$
U-Net-64 ( $256^3$ )	$2.11 \pm 0.10$	SEU-Net-64 ( $256^3$ )	$3.37 \pm 0.10$
U-Net-128 ( $256^3$ )	$4.39 \pm 0.25$	SEU-Net-128 ( $256^3$ )	$8.10 \pm 0.50$

ent network structures [29]. The reason for the accuracy improvement is that large input and model yield big context and learning capacity, respectively. We investigate the impact of large input on inference time by averaging three rounds of inferences in table 3. Using large input in the inference reduces the inference time significantly because it reduces the number of inference rounds. Results on liver and tumor segmentation task validate large input increases segmentation accuracy and reduces the inference time in table 4 and 5.

## 4 Conclusion

In this work, we try to investigate the impact of model size and input context size on two medical image segmentation tasks. To run large models and large input in the GPUs, we design a parallel U-Net with sequential modification based on an automated parallelism. Extensive results demonstrate that, 1) large model and input increases segmentation accuracy, 2) large input reduces inference time significantly. The Large deep networks with Automated Model Parallelism (LAMP) can be a useful tool for many medical image analysis tasks such as large image registration [30, 31], detection and neural architecture search.



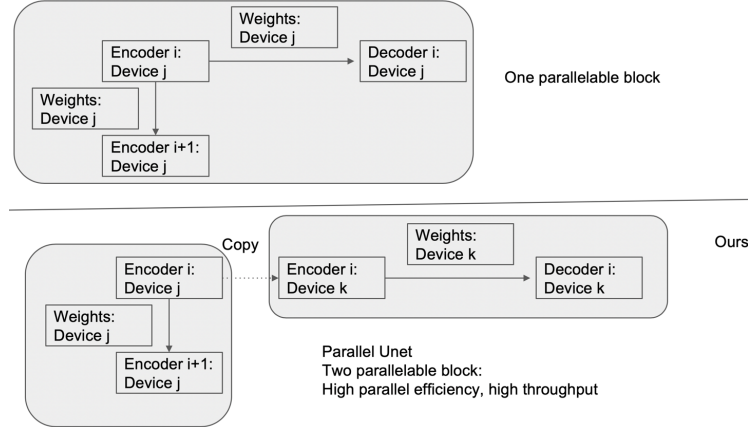
## References

1. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
2. Blumberg, S.B., Tanno, R., Kokkinos, I., Alexander, D.C.: Deeper image quality transfer: Training low-memory neural networks for 3d images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 118–125. Springer (2018)
3. Brügger, R., Baumgartner, C.F., Konukoglu, E.: A partially reversible u-net for memory-efficient volumetric image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 429–437. Springer (2019)
4. Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174 (2016)
5. Clark, K., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging* **26**(6), 1045–1057 (2013)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Gomez, A.N., Ren, M., Urtasun, R., Grosse, R.B.: The reversible residual network: Backpropagation without storing activations. In: Advances in neural information processing systems. pp. 2214–2224 (2017)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
10. Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., Lee, H., Ngiam, J., Le, Q.V., Wu, Y., et al.: Gpipe: Efficient training of giant neural networks using pipeline parallelism. In: Advances in Neural Information Processing Systems. pp. 103–112 (2019)
11. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. In: *Bildverarbeitung für die Medizin 2019*, pp. 22–22. Springer (2019)
12. Jesson, A., Guizard, N., Ghalehjegh, S.H., Goblot, D., Soudan, F., Chapados, N.: Cased: curriculum adaptive sampling for extreme data imbalance. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 639–646. Springer (2017)
13. Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B.: Boundary loss for highly unbalanced segmentation. In: International Conference on Medical Imaging with Deep Learning. pp. 285–296 (2019)
14. Lee, H., Jeong, M., Kim, C., Lim, S., Kim, I., Baek, W., Yoon, B.: torchgpipe, A GPipe implementation in PyTorch. <https://github.com/kakaobrain/torchgpipe> (2019)
15. Martens, J., Sutskever, I.: Training deep and recurrent networks with hessian-free optimization. In: *Neural networks: Tricks of the trade*, pp. 479–535. Springer (2012)
16. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al.: Mixed precision training. arXiv preprint arXiv:1710.03740 (2017)

17. Narayanan, D., Harlap, A., Phanishayee, A., Seshadri, V., Devanur, N.R., Ganger, G.R., Gibbons, P.B., Zaharia, M.: Pipedream: generalized pipeline parallelism for dnn training. In: *Proceedings of the 27th ACM Symposium on Operating Systems Principles*. pp. 1–15 (2019)
18. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. *OpenAI Blog* **1**(8), 9 (2019)
19. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimization towards training a trillion parameter models. *arXiv preprint arXiv:1910.02054* (2019)
20. Raudaschl, P.F., et al.: Evaluation of segmentation methods on head and neck ct: Auto-segmentation challenge 2015. *Medical physics* (2017)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241. Springer (2015)
22. Shazeer, N., Cheng, Y., Parmar, N., Tran, D., Vaswani, A., Koanantakool, P., Hawkins, P., Lee, H., Hong, M., Young, C., et al.: Mesh-tensorflow: Deep learning for supercomputers. In: *Advances in Neural Information Processing Systems*. pp. 10414–10423 (2018)
23. Shoyebi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., Catanzaro, B.: Megatron-lm: Training multi-billion parameter language models using gpu model parallelism. *arXiv preprint arXiv:1909.08053* (2019)
24. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019)
25. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 240–248. Springer (2017)
26. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging* **35**(5), 1299–1312 (2016)
27. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* **4**(2), 26–31 (2012)
28. Vallières, M., et al.: Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific reports* **7**(1), 10117 (2017)
29. Zhu, W., Huang, Y., Zeng, L., Chen, X., Liu, Y., Qian, Z., Du, N., Fan, W., Xie, X.: Anatomicnet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical physics* **46**(2), 576–589 (2019)
30. Zhu, W., Myronenko, A., Xu, Z., Li, W., Roth, H., Huang, Y., Milletari, F., Xu, D.: Neurereg: Neural registration and its application to image segmentation. In: *The IEEE Winter Conference on Applications of Computer Vision*. pp. 3617–3626 (2020)
31. Zhu, W., et al.: Neural multi-scale self-supervised registration for echocardiogram dense tracking. *arXiv preprint arXiv:1906.07357* (2019)
32. Zhuang, J., Dvornek, N.C., Li, X., Ventola, P., Duncan, J.S.: Invertible network for classification and biomarker selection for asd. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 700–708. Springer (2019)

## A Appendix: Design of LAMP

The figure 3 shows we reduce the dependency of long range skip-connection (Up) by separating it to two blocks (Bottom). Through the design of LAMP, the parallel U-Net achieves more parallel blocks, which lead to high throughput. We proof this in the next section.



**Fig. 3.** Up: The long range skip-connection hinders the parallelism in the U-Net. Bottom: We explicitly construct a variant of U-Net to remove the long range dependency in the U-Net. The parallel U-Net has higher parallel efficiency.

## B Appendix: Proof for High Throughput of LAMP

We demonstrate the parallel U-Net with LAMP has higher throughput in Fig. 4 and Fig. 5.

For  $t$  GPU time, Baseline: (time, block, batch, GPU id)

(1, 1, 1, 1);

(2, 2, 1, 2), (2, 1, 2, 1);

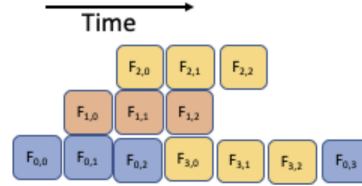
(3, 3, 1, 3), (3, 2, 2, 2), (3, 1, 3, 1)

(4, 4, 1, 1), (4, 3, 2, 3), (4, 2, 3, 2)

(5, 4, 2, 1), (5, 3, 3, 3)

(6, 4, 3, 1)

If  $t \gg 3$ , almost  $t/2$  batches for  $t$  GPU times.



**Fig. 4.** In the conventional U-Net based on pipeline parallelism, using three devices processes  $t/2$  batches in  $t$  device time.

For  $t$  GPU time,

Ours: (time, block, batch, GPU id)

(1, 1, 1, 1);

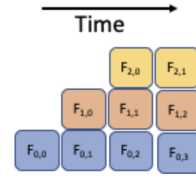
(2, 2, 1, 2), (2, 1, 2, 1);

(3, 3, 1, 3), (3, 2, 2, 2), (3, 1, 3, 1)

If  $t \gg 3$ , almost  $t$  batches for  $t$  GPU times.

Let us assume the GPU executing time is  $4/3$  of baseline, because of less blocks.

We can still run  $\frac{3}{4} * t$  batches. If we have  $K$  devices,  $K / (K+1) > 1/2$



**Fig. 5.** In the parallel U-Net based on LAMP parallelism, using three devices processes  $3t/4$  batches in  $t$  device time. Parallel U-Net based on LAMP has a higher throughput.