MedContext: Learning Contextual Cues for Efficient Volumetric Medical Segmentation

Hanan Gani[1], Muzammal Naseer[1], Fahad Khan[1,2], and Salman Khan[1,3]

[1] Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)
[2] Linköping University
[3] Australian National University
{hanan.ghani,muzammal.naseer,fahad.khan,salman.khan}@mbzuai.ac.ae

Abstract. Deep neural networks have significantly improved volumetric medical segmentation, but they generally require large-scale annotated data to achieve better performance, which can be expensive and prohibitive to obtain. To address this limitation, existing works typically perform transfer learning or design dedicated pretraining-finetuning stages to learn representative features. However, the mismatch between the source and target domain can make it challenging to learn optimal representation for volumetric data, while the multi-stage training demands higher compute as well as careful selection of stage-specific design choices. In contrast, we propose a universal training framework called MedContext that is architecture-agnostic and can be incorporated into any existing training framework for 3D medical segmentation. Our approach effectively learns self-supervised contextual cues jointly with the supervised voxel segmentation task without requiring large-scale annotated volumetric medical data or dedicated pretraining-finetuning stages. The proposed approach induces contextual knowledge in the network by learning to reconstruct the missing organ or parts of an organ in the output segmentation space. The effectiveness of MedContext is validated across multiple 3D medical datasets and four state-of-the-art model architectures. Our approach demonstrates consistent gains in segmentation performance across datasets and architectures even in few-shot scenarios. Our code is available at https://github.com/hananshafi/medcontext

Keywords: Volumetric medical segmentation · Masked image modeling · Knowledge distillation

1 Introduction

Deep neural networks have greatly improved volumetric medical segmentation. The convolutional encoder-decoder networks, U-NET [26,9], as well as the development of vision transformers [12], has led to hybrid architectures [20,4] with complementary strengths of self-attention and convolution for medical segmentation. Despite the architectural advances, deep neural networks generally require large-scale annotated data to achieve better performance. However, collecting

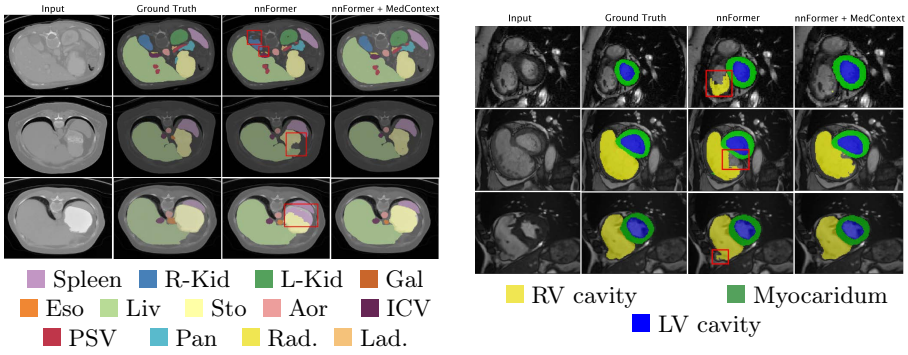

Fig. 1. Qualitative Comparison between the baseline nnFormer and our proposed MedContext integrated with nnFormer. The examples display different abdominal organs (Synapse) (**Left**) and regions of the heart (ACDC) (**Right**), with their labels in the legend below. The baseline nnFormer struggles to accurately segment the organs and heart regions, giving false segmentation results highlighted in red boxes. Best viewed zoomed in. Refer to supplementary material for additional qualitative comparisons.

and annotating medical images at a large scale can be expensive and prohibitive due to privacy concerns.

To deal with the data scarcity, weights learned on ImageNet [10] can be used to initialize the encoder, however, pre-training on 2D natural images may not capture the contextual information essential to understanding 3D medical images. Recent studies [15,17,28] explore self-supervised pre-training on extra auxiliary medical data, but this approach has two limitations: a) it involves a computationally expensive two-stage pre-training process on auxiliary data followed by fine-tuning on target data, and b) the success of fine-tuning depends on how well the auxiliary data distribution matches the target data. Moreover, there may not be a direct relationship between the self-supervised objectives and voxel-wise segmentation. Therefore jointly optimizing such self-supervised losses with 3D segmentation is non-trivial.

To address these limitations, we propose a generic training framework dubbed *MedContext* to learn self-supervised contextual cues jointly with supervised voxel segmentation without requiring large-scale annotated volumetric medical data. Our approach involves reconstructing masked organs or organ parts in the output segmentation space. This reconstruction aligns well with the voxel-wise prediction task, enabling joint optimization of both tasks. To further reduce the disparity between the two tasks, we deploy a student-teacher distillation strategy [25,5,19] to guide reconstruction from a slow-moving online teacher model which also helps avoid representation collapse. Predicting the representation of an input from a representation of another input leads to versatile visual representations [2]. MedContext encourages contextual learning within the model and allows it to learn local-global relationships between different input components. This leads to better segmentation of organ boundaries (see Fig. 1).

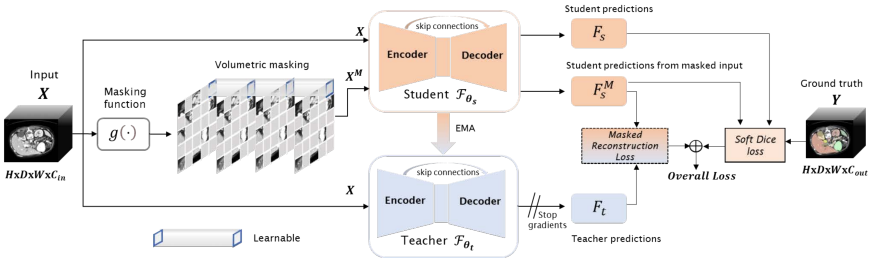

Fig. 2. *Overview of our MedContext approach:* The original 3D volume is masked and fed to the student model (top-row) along with the original input. The teacher model (bottom-row) is only fed with the original volume. The difference between the semantic voxelwise predictions for the masked and original inputs corresponding to the student and teacher networks respectively is minimized to guide the reconstruction of masked regions in the output segmentation space.

Our proposed approach is architecture-agnostic and can be incorporated into any training framework, making it universally applicable. We integrate our approach into three recent state-of-the-art medical 3D transformer based architectures: UNETR [16], SwinUNETR [15] and nnFormer [30]; and one CNN based 3D architecture PCRLv2 [31]. Using these architectures, we validate our approach across three medical imaging datasets: Multi-organ Synapse [21], ACDC [3] and BraTS [22,1]. Our evaluation reveals consistent performance improvements across all compared methods. In summary, our contributions are threefold: (1) We propose a universal training framework to jointly optimize supervised segmentation and self-supervised segmentation reconstruction via student-teacher knowledge distillation. (2) Our approach induces contextual knowledge in the model by learning to reconstruct the missing organ or organ parts in the output segmentation space. (3) We validate the effectiveness of our approach across multiple 3D medical datasets and state-of-the-art model architectures.

2 Methodology

2.1 Architecture

Our approach is complementary and can be applied to the existing encoder-decoder architectures designed for 3D medical image segmentation. As shown in Fig. 2 our design includes a student $\mathcal{F}_s$ and a teacher network $\mathcal{F}_t$ that operate on the input volume $\boldsymbol{X} \in \mathbb{R}^{H \times W \times D}$ and its masked version $\boldsymbol{X}^M \in \mathbb{R}^{H \times W \times D}$ generated using the masking function $g(.)$. Here, H, W, and D represent the height, width, and depth of the 3D volume, respectively. During the training phase, the input views are fed to the student-teacher framework as 3D patches, generating voxel-wise semantic logits for each input view. The student network is provided with both the masked ($\boldsymbol{X}^M$) and unmasked ($\boldsymbol{X}$) inputs, and the corresponding output voxel-wise semantic logits are denoted as $\boldsymbol{F}_s$ and $\boldsymbol{F}_s^M$,

respectively. On the other hand, the teacher network is provided with the original unmasked input $\boldsymbol{X}$ which outputs voxel-wise semantic logits denoted as $\boldsymbol{F}_t$. The feature map produced at intermediate layers of the 3D architecture has a shape of $\frac{H}{P_1} \times \frac{W}{P_2} \times \frac{D}{P_3} \times C$, where (P_1, P_2, P_3) is the resolution of each patch and C is the feature dimension. For each output prediction from the student, a supervised loss is computed using the ground truth label $\boldsymbol{Y}$, as shown in the figure. Additionally, a self-supervised objective is minimized between the masked student logits $\boldsymbol{F}_s^M$ and the teacher logits $\boldsymbol{F}_t$. Finally, both the supervised and self-supervised objectives are jointly optimized during single-stage training process.

2.2 Volumetric Masking Strategy

To model contextual relationships, we employ a masking technique on the patch tokens of the original input $\boldsymbol{X}$ to reconstruct missing parts in the segmentation space. We ensure mask consistency across the depth to prevent information leakage from neighboring cubes by applying the same mask to all subsequent slices in the volume as shown in Fig. 2. To generate a masked view $\boldsymbol{X}^M$, we randomly mask a certain fraction δ of the patch tokens. Following [11], the masked tokens are replaced with learnable tokens $\mathcal{H}_\xi$, such that,

$$\boldsymbol{X}^M = g(\boldsymbol{X}, \delta) = \boldsymbol{X} \circ (1 - \boldsymbol{I}_\delta) + \mathcal{H}_\xi \circ \boldsymbol{I}_\delta, \tag{1}$$

where $\boldsymbol{I}_\delta$ is a binary mask generated according to a Bernoulli distribution using $g(.)$, i.e., $\boldsymbol{I}_\delta \sim \text{Bernoulli}(\delta)$ and $\circ$ denotes the element-wise product.

2.3 Voxel-wise Segmentation Reconstruction

We utilize masked input to reconstruct segmentation maps, facilitating the learning of contextual semantic relationships. To achieve this, we employ a student-teacher strategy where teacher weights are updated by a moving average of the student weights. Leveraging cumulative knowledge from prior weight updates enhances masked view reconstruction and induces enriched contextual cues. The teacher network provides soft semantic targets, guiding student network training. Both the student model $\mathcal{F}_s$ and teacher model $\mathcal{F}_t$ begin with the same randomly initialized weight parameters. The student network processes both original and masked inputs, while the teacher network only receives the original non-masked input. The networks generate voxel-wise semantic logits, represented by $\{\boldsymbol{F}_s^M, \boldsymbol{F}_s\}$ and $\boldsymbol{F}_t$ respectively. Subsequently, we reconstruct semantic voxel-wise logits of the masked input from the student model, guided by two supervised signals: *supervision through knowledge distillation* and *ground truth labels*.

Reconstruction through Knowledge Distillation: A self-supervised distillation loss (Eq. 2) is used to guide the training of the student network to encourage modeling the contextual consistency. It minimizes the difference between the voxel-wise logits $\boldsymbol{F}_t$ generated by the teacher network given the original input $\boldsymbol{X}$ and the voxel-wise logits $\boldsymbol{F}_s^M$ produced by the student network using

the masked input $\boldsymbol{X}^M$. The objective function, referred to as Consistency Loss (CL), is denoted as $\mathcal{L}_c(\boldsymbol{F}_s^M, \boldsymbol{F}_t)$ and is expressed as,

$$\mathcal{L}_c(\boldsymbol{F}_s^M, \boldsymbol{F}_t) = \frac{\| \boldsymbol{F}_s^M - \boldsymbol{F}_t \|_2^2}{\| \boldsymbol{F}_t \|_2^2}. \tag{2}$$

Reconstruction through Ground truth Labels: The voxel-wise semantic logits $\boldsymbol{F}_s^M$ output by the student for $\boldsymbol{X}^M$ are further reconstructed using the ground truth labels. This is achieved by minimizing the soft dice loss [23] using the ground truth labels $\boldsymbol{Y}$. The general expression for Dice-CE Loss for some arbitrary output prediction $\boldsymbol{F}$ is given as,

$$\mathcal{L}_{Dice-CE}(\boldsymbol{Y}, \boldsymbol{F}) = 1 - \sum_{c=1}^{C} \left(\frac{2 * \sum_{v=1}^{V} \boldsymbol{Y}_{v,c} \cdot \boldsymbol{F}_{v,c}}{\sum_{v=1}^{V} \boldsymbol{Y}_{v,c}^2 + \sum_{v=1}^{V} \boldsymbol{F}_{v,c}^2} + \sum_{v=1}^{V} \boldsymbol{Y}_{v,c} \log \boldsymbol{F}_{v,c} \right), \tag{3}$$

where, C denotes the number of classes; V denotes the number of voxels; $\boldsymbol{Y}_{v,i}$ and $\boldsymbol{F}_{v,i}$ denote the ground truths and output probabilities for class i at voxel v, respectively. In our case, the supervised reconstruction objective is calculated using above Dice-CE loss between the ground truth label $\boldsymbol{Y}$ and voxel-wise semantic logits $\boldsymbol{F}_s^M$ and is denoted as $\mathcal{L}_{Dice-CE}(\boldsymbol{Y}, \boldsymbol{F}_s^M)$ and referred to as Masked Student Loss (MSL). Both CL and MSL encourage the network to capture intricate relationships between various organs.

2.4 Supervised Voxel-wise Segmentation

Our primary task of supervised voxel-wise segmentation takes place in conjunction with the voxel-wise segmentation reconstruction as discussed above. For the supervised voxel-wise segmentation, we optimize the predictions of the of the student network $\boldsymbol{F}_s$ on $\boldsymbol{X}$ through the supervision of the ground truth labels $\boldsymbol{Y}$ using Soft Dice Loss (Eq. 3) denoted by the objective $\mathcal{L}_{Dice-CE}(\boldsymbol{Y}, \boldsymbol{F}_s)$.

2.5 Overall Multi-task Objective

Our framework leverages a combination of supervised and self-supervised losses for optimization, synergistically reinforcing each other to offer complementary advantages. The overall loss objective $\mathcal{L}$ is defined as,

$$\mathcal{L} = \mathcal{L}_{Dice-CE}(\boldsymbol{Y}, \boldsymbol{F}_s) + \mathcal{L}_{Dice-CE}(\boldsymbol{Y}, \boldsymbol{F}_s^M) + \beta \mathcal{L}_c(\boldsymbol{F}_s^M, \boldsymbol{F}_t), \tag{4}$$

where the hyperparameter β controls the contribution of self-supervised consistency loss during optimization.

2.6 Optimization strategy

Following a typical student-teacher optimization strategy as ultilized by [13,5], the gradient of the total loss is backpropagated through the student network

and parameters are updated as: $\Theta \leftarrow \Theta - \alpha \cdot \nabla_{\Theta}(\mathcal{L})$, where Θ represents the joint parameters of student network (θ_s) and learnable mask embeddings (ξ) i.e. $\Theta = \{\theta_s; \xi\}$. The teacher network is updated via exponential moving average (EMA) of the weights of the student network using: $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$, where θ_t denote the parameters of teacher and λ follows the cosine schedule from 0.996 to 1 during training. The gradient step through the student network comprises of the contributions from both the supervised and self-supervised objectives, thereby aiding in the reconstruction of the masked input by updating the differentiable volumetric embeddings associated with the masked regions.

3 Experiments

Datasets: We evaluate on three volumetric medical datasets. **Synapse BTCV**: The synapse BTCV dataset [21] for multi-organ CT Segmentation, includes abdominal CT scans of 30 subjects. We adopt the dataset split of [6] with 18 train and 12 test samples. We evaluate the performance on eight abdominal organs. **ACDC:** The ACDC dataset [3] is a collection of cardiac MRI images and associated segmentation annotations for the right ventricle (RV), left ventricle (LV), and myocardium (MYO) of 100 patients. We split the dataset into 80 training and 20 testing samples following [30]. **BraTS:** We use two versions of BraTS dataset: BraTS17 [22] and BraTS21 [1]. For *UNETR, SwinUNTER* and *PCRLv2* we report results on the BraTS21 dataset to be consistent with the baseline. The BraTS21 dataset includes 1251 subjects with annotations for three sub-regions: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). Following [15], we train on 1000 subjects and test on 251 subjects. For *nnFormer*, we use BraTS17 dataset comprising of 484 MRI images. Following [30], we train on 387 samples and test on 73 cases. **Evaluation Metrics:** To evaluate the models' performance, we utilize two metrics: the Dice Similarity Score (DSC) and the 95% Hausdorff Distance (HD95). **Training and Implementation details:** Our approach utilizes Pytorch version 1.10.1 in conjunction with MONAI libraries [24] for implementation. Specifically, we use an input size of $128 \times 128 \times 64$ for all datasets when training with *nnFormer*, and $96 \times 96 \times 96$ for *UNETR*, *SwinUNTER* and *PCRLv2*. All models are trained on a single A100 40GB GPU.

3.1 Comparison with state-of-the-art Baselines

Synapse BTCV Dataset: Table 1 shows the results on the synapse multi-organ dataset. UNETR with our approach achieves 2.5% higher Dice Score (81.13%) than the baseline (78.76%), and over 1% reduction in HD95 score. With hierarchical SwinUNETR, Dice Score increases by >1% (80.66% to 82.00%) and HD95 improves. Similar trend is observed with nnFormer. **ACDC Dataset**: Table 2 presents results on the larger ACDC dataset. UNETR with our approach outperforms the baseline by about 4% in Dice score (80.60% vs. 76.67%). SwinUNETR shows over 2.5% Dice score improvement, and nnFormer achieves a Dice score of 90.73% compared to the baseline's 90.50%. Per organ dice scores are also

Table 1. Abdominal multi-organ Synapse: Our MedContext consistently improves the segmentation performance of all organs across different models. We observe significant improvements in HD95 along with the dice score (DSC). Best results in bold.

Models	MedContext	Spleen	Right Kidney	Left Kidney	Gallbladder	Liver	Stomach	Aorta	Pancreas	Average	
										HD95 ↓	DSC ↑
UNETR	✗	89.64	83.02	84.86	63.06	95.58	73.06	87.47	53.40	11.04	78.76
	✓	90.73	83.36	86.03	67.94	95.59	78.62	87.30	59.51	**9.44**	**81.13**
Swin-UNETR	✗	86.33	80.63	84.07	67.24	94.98	74.97	90.53	66.49	20.32	80.66
	✓	91.45	80.80	84.85	67.70	94.60	76.20	90.88	67.74	**14.45**	**82.00**
nnFormer	✗	90.51	86.25	86.57	70.17	96.84	86.83	92.04	83.35	10.63	86.57
	✓	95.97	87.05	87.63	72.87	96.43	84.57	91.85	82.40	**8.29**	**87.35**

Table 2. ACDC: We report DSC on RV, LV and MYO.

Models	MedContext	RV	Myo	LV	Average
UNETR	✗	77.81	72.74	79.46	76.67
	✓	84.77	75.82	81.21	**80.60**
SwinUNETR	✗	83.47	75.54	83.09	80.70
	✓	84.79	79.17	86.15	**83.38**
nnFormer	✗	91.18	86.24	94.07	90.50
	✓	92.14	86.52	93.52	**90.73**

Table 3. BraTS: We report DSC on 3 brain tumour types

Models	MedContext	WT	ET	TC	Average
UNETR	✗	87.35	90.88	84.29	87.50
	✓	87.43	91.45	85.23	**88.04**
SwinUNETR	✗	90.36	91.72	86.24	89.44
	✓	90.57	92.30	86.64	**89.83**
nnFormer	✗	80.80	58.86	77.42	72.36
	✓	81.00	59.87	77.45	**72.78**

Table 4. Few-shot settings (5 train samples).

Models	MedContext	Synapse	ACDC
UNETR	✗	53.83	18.53
	✓	**56.25**	**28.63**
SwinUNETR	✗	54.13	32.62
	✓	**61.15**	**35.80**
nnFormer	✗	67.90	52.23
	✓	**70.96**	**58.05**

Table 5. MedContext vs. pretraining-finetuning [8] methods. DSC (%) on Synapse dataset with UNETR architecture. Best viewed zoomed in.

Method	Pretrain	Spleen	RKid	LKid	Gall	Eso	Liv	Sto	Aorta	IVC	Veins	Pan	RAG	LAG	Avg
Baseline	✗	89.0	89.2	87.7	47.6	48.9	94.4	74.7	82.0	77.3	61.7	64.4	56.6	46.9	70.8
SimCLR	✓	91.1	91.3	89.7	48.7	50.0	96.6	76.5	83.9	79.1	63.2	65.9	57.9	48.1	72.4
MAE	✓	94.8	95.0	93.4	50.6	52.1	98.6	79.7	87.4	82.4	65.9	68.6	60.5	50.1	75.3
SimMIM	✓	95.2	95.4	93.7	51.9	52.3	98.7	79.9	87.7	82.6	66.0	68.9	60.7	51.2	75.7
MedContext	✗	93.8	93.7	93.6	54.9	72.6	96.6	80.3	89.9	83.3	72.9	73.9	64.4	65.3	**79.6**

Table 6. Improving PCRLv2 with our proposed MedContext without pretraining across three datasets. We report Average Dice scores (%).

Method	Pretrain	Brats21	ACDC	Synapse
PCRLv2	✓	79.90	78.53	64.00
PCRLv2 + MedContext	✗	**82.03**	**82.57**	**72.30**

higher with our approach in each case. **BraTS Dataset**: Table 3 demonstrates that UNETR yields a 0.54% increase in the overall DSC compared to baseline. SwinUNTER achieves a DSC of 89.83%, surpassing the baseline DSC 89.44%. Additionally, nnFormer exhibits an improvement in the overall DSC (72.78%) compared to the baseline DSC (72.36%). Overall we show that our approach achieves gains even on larger datasets such as BraTS, but has more pronounced improvement for the low-data setups. See further results in Supplementary.

3.2 Few-shot performance

We validate our approach in a few-shot scenario in Table 4, comparing its performance with baselines in a 5-shot setting on synapse BTCV and ACDC datasets using three model architectures. Specifically, on synapse, our approach yields a 3-10% increase in Dice score across all cases, indicating substantial segmentation accuracy improvement. Similarly, on the larger ACDC dataset, our approach consistently achieves higher Dice scores compared to baselines, highlighting its potential for enhancing segmentation accuracy in situations with limited annotated data and supporting data-efficient training.

Table 7. Effect of each loss component. We report avg dice score (%).

MSL	CL	Average Dice Score	
		UNETR	SwinUNETR
✓	✗	78.69	81.03
✗	✓	79.46	81.25
✓	✓	**80.32**	**81.70**

Table 8. Effect of masking ratio. We report average DSC (%).

Masking ratio	Average Dice Score	
	UNETR	SwinUNETR
30%	79.54	80.92
40%	**80.47**	**82.00**
50%	80.00	81.03
60%	80.20	81.70
80%	79.90	81.27

Table 9. Effect of knowledge distillation for leveraging contextual cues.

Models	Student-Teacher	Average DSC
UNETR	✗	79.60
	✓	81.13
SwinUNETR	✗	80.83
	✓	**82.03**
nnFormer	✗	86.85
	✓	**87.36**

3.3 Comaprison with Pretraining-Finetuning Baselines

We demonstrate the effectiveness of MedContext by comparing its performance (DSC) with existing pretraining-finetuning methods in Table 5. The baseline [8] utilizes improved weight initialization through pretraining on a large dataset [14], incorporating state-of-the-art self-supervised methods [7,29,18], and then fine-tunes on the target dataset. In contrast, MedContext directly learns contextual cues from the small target dataset, outperforming methods using the pretraining-finetuning paradigm. When integrated into PCRLv2 [31], a 3D CNN architecture pretrained in a self-supervised manner on [27], MedContext consistently enhances the performance (Avg. DSC) of PCRLv2 without pretraining as seen in Table 6, affirming its versatility applicable to various CNN architectures.

3.4 Ablation Studies

Effect of different losses: Our proposed method incorporates multiple supervised and self-supervised losses during training as elaborated in Sections 2.3 and 2.4. We perform an ablative analysis on the synapse dataset, focusing on the Masked Student Loss (MSL) and Consistency Loss (CL) in Table 7. Eliminating either loss component leads to a decrease in DSC, underscoring the mutual synergy between these losses in inducing contextual cues for effective 3D medical segmentation. **Effect of Student-Teacher framework:** Using a single model for both original and masked input using supervised loss may not effectively capture contextual relationships as it overlooks knowledge acquired during previous weight updates. To address this, we adopt a student-teacher framework, leveraging information from past updates. Table 9 illustrates our claim with empirical evidence on the synapse dataset, showing a consistent performance drop without the student-teacher framework. **Effect of masking ratio:** Our method encourages learning contextual cues by reconstructing masked regions in the segmentation space. The fraction of patches to be masked for reconstruction may influence the model's performance. Our approach produces gains on all masking ratios, however, our analysis in Table 8 reveals a 40% masking ratio to be optimal for learning contextual cues.

4 Conclusion

In this paper, we propose a universal training framework called *MedContext* which effectively learns self-supervised contextual cues jointly with the supervised voxel segmentation task without requiring large-scale annotated volumetric medical data. Our proposed approach employs a student-teacher distillation strategy to reconstruct missing parts in the output segmentation space. Through extensive experimentation, our approach demonstrates complementary benefits to existing 3D medical segmentation architectures in both conventional and few-shot settings without pretraining on large-scale datasets. Moreover, the plug-and-play design of our approach allows for its easy integration into any architectural design.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baid, U., Ghodasara, S., Mohan, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
2. Bardes, A., Garrido, Q., Ponce, J., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint (2024)
3. Bernard, O., Lalande, A., Zotti, C.e.a.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (2018)
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision Workshops (2022)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
8. Chen, Z., Agarwal, D., Aggarwal, K., Safta, W., Balan, M.M., Brown, K.: Masked image modeling advances 3d medical image analysis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1970–1980 (January 2023)
9. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016. pp. 424–432. Springer (2016)

10. 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 (2009)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. preprint arXiv:1810.04805 (2018)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
14. Harmon, S.A., Sanford, T., Xu, S., et al.: Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. Nature Communications **11** (2020)
15. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop (2022)
16. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
17. Hatamizadeh, A., Xu, Z., Yang, D., Li, W., Roth, H., Xu, D.: Unetformer: A unified vision transformer model and pre-training framework for 3d medical image segmentation. arXiv preprint arXiv:2204.00631 (2022)
18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
19. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
20. Karimi, D., Vasylechko, S.D., Gholipour, A.: Convolution-free medical image segmentation using transformers. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021. pp. 78–88. Springer (2021)
21. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge (2015)
22. Menze, B.H., Jakab, A., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
23. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision (3DV) (2016)
24. Project-MONAI: Medical open network for ai. https://github.com/Project-MONAI/MONAI (2020)
25. Richemond, P.H., Grill, J.B., Altché, F., et al.: Byol works even without batch statistics. arXiv preprint arXiv:2010.10241 (2020)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)

27. Setio, A.A.A., Traverso, A., et. al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge. Medical Image Analysis **42**, 1–13 (2017)
28. Xie, Y., Zhang, J., Xia, Y., Wu, Q.: Unified 2d and 3d pre-training for medical image classification and segmentation. arXiv preprint arXiv:2112.09356 (2021)
29. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)
30. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)
31. Zhou, H.Y., Lu, C., Chen, C., Yang, S., Yu, Y.: Pcrlv2: A unified visual information preservation framework for self-supervised pre-training in medical image analysis. arXiv preprint arXiv:2301.00772 (2023)