Accurate Hippocampus Segmentation Based on Self-supervised Learning with Fewer Labeled Data

Kassymzhomart Kunanbayev(✉), Donggon Jang, Woojin Jeong, Nahyun Kim, and Dae-Shik Kim

KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, South Korea
{kkassymzhomart,jdg900,woojin05,nhkim21,daeshik}@kaist.ac.kr

Abstract. Brain MRI-based hippocampus segmentation is considered as an important biomedical method for prevention, early detection, and accurate diagnosis of neurodegenerative disorders like Alzheimer's disease. The recent need for developing accurate as well as robust systems has led to breakthroughs making advantage of deep learning, but requiring significant amounts of labeled data, which, in turn, is costly and hardly obtainable. In this work, we try to address this issue by introducing self-supervised learning for hippocampus segmentation. We devise a new framework, based on the widely known method of Jigsaw puzzle reassembly, in which we first pre-train using one of the unlabeled MRI datasets, and then perform a downstream segmentation training with other labeled datasets. As a result, we found our method to capture local-level features for better learning of anatomical information pertaining to brain MRI images. Experiments with downstream segmentation training show considerable performance gains with self-supervised pre-training over supervised training when compared over multiple label fractions.

Keywords: Hippocampus · Segmentation · Self-supervised learning · Jigsaw puzzle reassembly

1 Introduction

There have been numerous studies linking the occurrence of Alzheimer's disease (AD) and the hippocampal part of the brain [14,21]. More specifically, the changes in the volume and structure of the hippocampus are associated with the level of the progression of AD [9,14]. In this regard, clinical analysis and timely diagnosis of the hippocampus are crucial for prevention and treatment. One of the most widely applied methods is hippocampus segmentation of brain magnetic resonance imaging (MRI), which can yield the necessary analytical information regarding the size and morphology of the hippocampal part of the brain. However, given its small size as well as the uniformity of MRI images, along with the importance of precise segmentation, the job of hippocampus segmentation turns out to be costly, time-consuming and requires highly qualified expertise to perform the task.

A. Abdulkadir et al. (Eds.): MLCN 2022, LNCS 13596, pp. 42–51, 2022.
https://doi.org/10.1007/978-3-031-17899-3_5

In this regard, the research community focused on devising more sophisticated strategies that could introduce methods with higher accuracy and robustness. To this end, machine and deep learning-based techniques have recently succeeded in automating and improving the task of hippocampus segmentation [2,6,16,20]. However, despite the success of these methods in automating and speeding up the task, achieved frameworks remain to require large amounts of MRI images with corresponding label masks, which are expensive and laborious to collect.

To address this issue, in computer vision, self-supervised learning methods have been introduced that avoid large amounts of labeled data by learning representations from unlabeled data via pretext task strategies [5,8]. In the medical image domain as well, there have been various works incorporating self-supervised learning [7,15]. Perhaps, one of the notable self-supervised learning methods is the widely known Jigsaw puzzle reassembly, proposed by [17], in which a network takes tiles of an image one at a time as input and solves the puzzle by predicting a correct spatial arrangement, thus learning the feature mapping of object parts. Experiments suggested that pre-training and then transferring weights for retraining on a downstream task outperforms supervised learning [17]. Taleb *et al.* [24] introduced solving multimodal Jigsaw puzzle for medical imaging by incorporating the Sinkhorn operator and exploiting synthetic images during pre-training. Navarro *et al.* [15] analyzed the self-supervised learning for robustness and generalizability in medical imaging. Although these methods utilize a Jigsaw puzzle, there are no specific previous works on hippocampus segmentation. Moreover, previous works usually tend to use the same datasets both for pre-training and downstream training, while in real-world scenarios the target dataset may not be large enough for pre-training.

In this regard, we further devise a new self-supervised framework for the task of hippocampus segmentation by adopting the Jigsaw puzzle reassembly problem. We selected this method because it explicitly limits the context of the network processing as it takes one tile at a time, hence, this arrangement should be useful to learn the anatomical parts of the brain and especially the features related to the hippocampal part, given its relatively small size in the brain. We first pre-train the model on one of the unlabeled brain MRI datasets and then re-train on a downstream segmentation task with other labeled datasets, by experimenting over various labeled data fractions (from 100% to 10%). Both quantitative and qualitative results show that pre-trained initialization leads to considerable performance gains in hippocampus segmentation.

2 Method

2.1 How Jigsaw Puzzle is Solved

The original implementation of Jigsaw Puzzle reassembly [17] used the context-free architecture, by building a siamese-ennead convolutional network with shared weights based on AlexNet [13]. The image tiles are first randomly permuted so that image patches are reordered, and fed to the network one at a

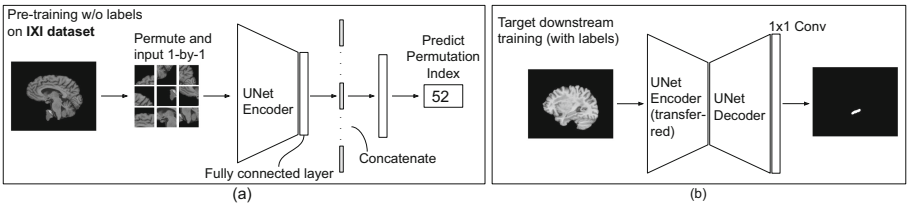

Fig. 1. Self-supervised learning-based (a) pre-training and (b) downstream training framework for hippocampus segmentation.

time. This architecture computes features for each tile separately and then concatenates them to feed as an input to the final fully connected layer (see Fig. 1). Afterwards, the network needs to predict the index of the permutation. The output of the network can be regarded in terms of conditional probability density as follows:

$$p(S|I_1, I_2, ..., I_9) = p(S|F_1, F_2, ..., F_9) \prod_{i=1}^{9} p(F_i|I_i), \tag{1}$$

where S is a permutation configuration, I_i is the i-th image tile, while F_i is the i-th corresponding feature representation after the final fully connected layer. If S is written as a list of positions $S = (L_1, .., L_9)$, then conditional probability distribution would be decomposed into independent terms:

$$p(L_1, L_2, ..., L_9|F_1, F_2, ..., F_9) = \prod_{i=1}^{9} p(L_i|F_i), \tag{2}$$

which implies that each tile position L_i is defined by corresponding F_i [17].

Moreover, one important contribution of Jigsaw puzzle reassembly is how it ensures learning correct representations that will be useful in leveraging the performance during the target downstream training, not just in solving the pretext Jigsaw puzzle task. This is because the network may be prone to learning the latter by following simpler solutions, called *shortcuts* [17]. In the following subsection, we discuss our framework for hippocampus segmentation and how we incorporated the idea of preventing shortcuts during pre-training.

2.2 Framework for Hippocampus Segmentation

Our training framework generally follows the original Jigsaw implementation [17], but with certain adjustments for medical imaging. For example, instead of the AlexNet network, we use the UNet encoder-decoder network [19] for its popularity and remarkable performance in medical imaging, as well as the convenient design. Instead of training the entire network for Jigsaw puzzle reassembly, which could be both time-consuming and computationally expensive, we pre-train only the encoder part and use its weights as initialization for downstream segmentation training. The framework flow is visualized in Fig. 1b.

As for Jigsaw puzzle reassembly, we divide the MRI image into a 3×3 grid to obtain 9 tiles. To further avoid shortcuts, the tiles are randomly cropped to a smaller size so that the model avoids solving the problem by learning edge-related features instead of object positions, so cropping will ensure random shifts in edges [17].

The UNet encoder network followed by fully connected linear layers needs to predict the permutation index. Note that for 9 tiles, there are $9! = 362880$ permutations possible. However, as in [17], to ensure that tiles are shuffled well enough, we selected 1000 permutations based on the Hamming distance. In this way, the network will predict one of the indexes from these 1000 permutations.

Downstream segmentation training is performed by initializing the encoder with pre-trained weights and randomly initializing the decoder. By pre-training only the encoder we ensure that the encoder learns global anatomical features needed to localize the hippocampus, while the decoder will be trained along with encoder weights for accurate hippocampus segmentation. Additionally, 1×1 convolutional layer is added as the last layer to facilitate the segmentation. More details can be seen in Fig. 1.

2.3 Datasets

IXI Dataset. For pre-training, we obtained T1-weighted images of a widely known IXI dataset (https://brain-development.org/ixi-dataset/) which was collected from three different hospitals: Hammersmith (Philips 3T scanner), Guy's (Philips 1.5T scanner), and Institute of Psychiatry (GE 1.5T scanner). There are a total of 579 MRI images, each having 150 slices sized 250×250. To improve the training outcome, the brain MRI volumes were preprocessed in the following order: (i) brain extraction was applied using the Brain Extraction Tool [23] (available as a part of the FMRIB Software Library) in order to remove non-brain areas that could affect the following pre-processing steps; (ii) the intensity inhomogeneity was applied using the N3 package of the MINC toolkit [22]; (iii) min-max normalization was performed volume-wise to normalize the values.

EADC-ADNI HarP Dataset. For downstream training experiments, we utilized publicly available hippocampus segmentation HarP dataset that was collected as a part of the EADC-ADNI Harmonized Protocol project [1,3,10,12,18] from the Alzheimer's disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu). The dataset contains 135 T1-weighted MRI volumes [18], consisting of 197 slices, each with a size of 189×233, and their released segmentation masks [4]. The same pre-processing steps as in the IXI dataset were applied to this dataset.

Decathlon Dataset. Additionally, we used a hippocampus segmentation dataset from the Medical Segmentation Decathlon challenge [11]. The dataset includes 265 training and 130 test volumes, but we only utilized the training set due to the unavailability of label masks for the test set. The volumes contain

cropped parts of the hippocampal area each with various sizes, hence we only selected images with a size of 32×32. No pre-processing was applied.

2.4 Experimental Setup

Pre-training. The pre-training phase was performed using the IXI dataset. To learn the relevant feature representations from the images that include the hippocampus, we discard the MRI images (slices) that do not contain substantial and meaningful brain information and utilize only images that vividly contain the brain anatomical parts. We centrally crop a 225×225 pixel image from the given MRI image and divide it into 9 tiles, each with a size of 75×75. We further crop it to a size of 64×64 to prevent shortcuts and resize the cropped tiles back to 75×75. The training was performed for 70 epochs with a batch size of 64. The stochastic gradient descent optimizer was used as in [17] but with an initial learning rate of 0.001 and a decay rate of 0.1 every 30 epochs.

Downstream Segmentation Training. The dataset split was done subject-wise, so the segmentation training was performed on 90% of volumes of each labeled dataset, while the remaining 10% is set out as a test set for final testing. In order to perform in-depth analysis of the experiments, we conduct comparisons among three training settings: **finetuning**, **linear**, and **random initialization** (baseline). The **finetuning** implies using pre-trained weights of the encoder and training it along with the decoder, while **linear** indicates freezing of the pre-trained weights of the encoder and training only the decoder, and finally, **random initialization** indicates randomly initializing both the encoder and decoder. In addition, to evaluate the performance of the method on both large and small-sized datasets, we performed experiments on 10%, 20%, 50%, and 100% label fractions of the data and reported the results. The segmentation task training was performed for 70 epochs and was repeated 3 times to report the average. The batch size was also selected 64. The quantitative results were evaluated using the Dice coefficient. The Adam optimizer was used with an initial learning rate of 0.001 and a decay rate of 0.1 every 30 epochs.

Network Details. The UNet implementation that we used follows the structure of the original implementation [19], so its encoder part contains five double convolutional layers with 3×3 kernels, as well as a batch normalization and a ReLU activation after each convolutional layer. The max pooling operator with a stride of 2 is used between the double convolutional layers. Thus, the total number of trainable parameters in the UNet encoder is ~ 9.4 million. During the pretraining, there are two fully connected layers following the UNet encoder with output sizes of 1024 and 4096, each of which is followed with a ReLU activation and a dropout layer with a rate of 0.5. There is also the final classification layer with an output of 1000. After the pre-training, these layers are discarded, and only the weights of the encoder are transferred for further re-training. In the decoder part, there are four double convolutional layers but with in-between

Table 1. Hippocampus segmentation results (Dice coefficient, %) on **the HarP** test set. The downstream segmentation training was conducted with different label fractions of the training data set. The **bold** numbers indicate the best results.

	Label fractions			
	100%	50%	20%	10%
Linear (decoder training, ours)	**99.94**	**99.68**	88.50	89.32
Fine-tuning (ours)	99.93	86.00	**97.44**	**93.11**
Random initialization	94.45	94.09	97.22	91.60

Table 2. Hippocampus segmentation results (Dice coefficient, %) on **the Decathlon** test set. The downstream segmentation training was conducted with different label fractions of the training data set. The **bold** numbers indicate the best results.

	Label fractions			
	100%	50%	20%	10%
Linear (decoder training, ours)	**92.37**	90.90	81.84	63.33
Fine-tuning (ours)	87.18	**92.16**	**91.97**	**88.55**
Random initialization	76.60	88.89	64.47	68.50

upsampling operators with a scale factor of 2 and a bilinear transformation algorithm. After the last double convolutional layer of the decoder, to facilitate the segmentation, there is a final 1×1 convolutional layer with an output sigmoid activation.

The training of the framework was conducted on GPU servers with NVIDIA Titan RTX (24GB) and NVIDIA RTX A6000 (48GB). Python-based PyTorch deep learning framework was used for implementation. The implementation code is available at the following link: https://github.com/qasymjomart/ssl_jigsaw_hipposeg.

3 Results

3.1 Quantitative Results

Tables 1 and 2 illustrate the quantitative results comparing the fine-tuning and linear settings with the random initialization on the HarP and Decathlon datasets, respectively. The test results suggest the superior performance of models pre-trained via the proposed method over all label fractions. On the HarP dataset, the linear setting consistently resulted in the highest segmentation accuracy in high label fractions, while the fine-tuning setting demonstrated a higher performance in the case of lower label fractions. These observations explain how pre-trained weights contribute to the overall performance over various amounts

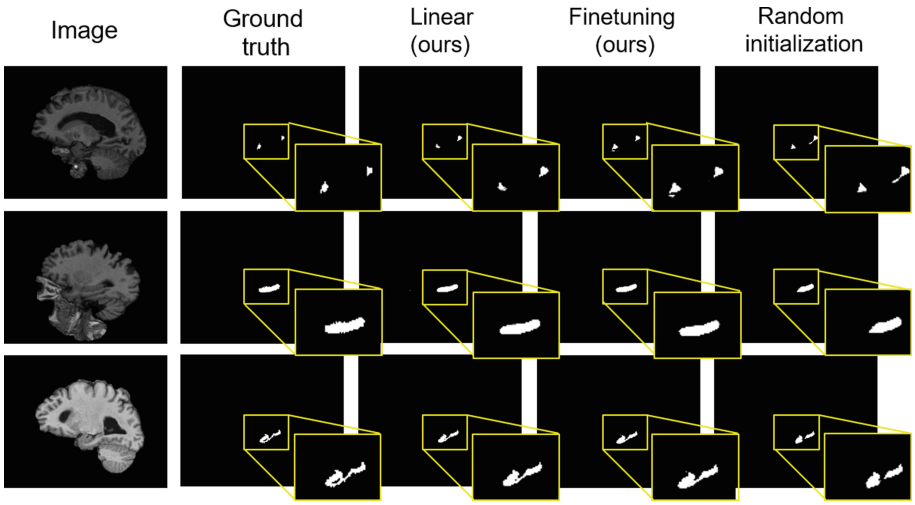

Fig. 2. Randomly selected qualitative results on **the HarP** test set. The results are from the downstream segmentation training with a label fraction of 10%.

of labeled data. Retraining using the pre-trained weights through the fine-tuning setting with fewer labeled data led to better alignment of the encoder and decoder weights in learning segmentation.

Similarly, the experiments on the Decathlon dataset yielded superior performance, where the fine-tuning allowed to achieve considerable performance gains as fewer data is utilized in downstream training. Large differences in accuracy of the 20% and 10% label fractions demonstrate that the fine-tuning setting turned out to be effective in leveraging the segmentation performance.

3.2 Qualitative Results

Qualitative results in Figs. 2 and 3 depict some of the randomly selected MRI brain images with corresponding ground truth and predicted masks. The masks predicted by our method on the HarP dataset exhibited significant similarity with the ground truth. In the first and third rows, the randomly initialized model overpredicted, while in the second row, it underpredicted the hippocampus area.

Similarly, Fig. 3 illustrates clear comparisons in the case of the Decathlon dataset. In all rows, the linear and random initialization demonstrated different predicted shapes and areas, meanwhile, the fine-tuning resulted in more resembling shapes and areas. These qualitative results agree with the quantitative results and suggest that the pre-training via the proposed method provides better capability in predicting the shape and edges of the hippocampus.

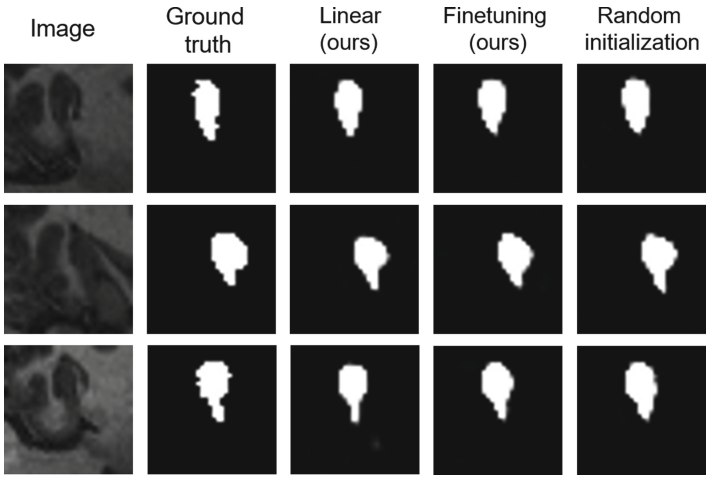

Fig. 3. Randomly selected qualitative results on **the Decathlon** test set. The results are from the downstream segmentation training with a label fraction of 10%.

4 Discussion and Conclusion

In general, results from both datasets indicate that the fine-tuning setting results in better segmentation performance, especially when fewer labeled data is utilized. This is particularly notable when only 10% of data is used for downstream training (see Tables 1 and 2). It is also important to note that for downstream training, we made use of all MRI slices, including those without clear brain anatomical features or even hippocampus, implying empty masks. This imbalance in data may lead to a failure in capturing brain-related features during downstream training. Hence, we conjecture that the effect of pre-trained weights may be better preserved in lower label fractions since they tend to have shorter training. Additionally, the linear settings showed that the pre-trained representations are useful in segmentation. More analysis, including cross-validation, is needed to further address this issue.

Developing accurate as well as robust systems for hippocampus segmentation is a prominent issue for early diagnosis and treatment of AD. In this work, a novel framework that may enhance the overall performance of such systems with fewer amounts of labeled data has been devised. The future research will focus on more in-depth analysis and further development of the framework using other state-of-the-art self-supervised learning methods, and comparing it with other hippocampus segmentation baselines.

Acknowledgements. This work was supported by the Engineering Research Center of Excellence (ERC) Program supported by National Research Foundation (NRF), Korean Ministry of Science & ICT (MSIT) (Grant No. NRF-2017R1A5A1014708).

References

1. Apostolova, L.G., et al.: Relationship between hippocampal atrophy and neuropathology markers: a 7t MRI validation study of the EADC-ADNI harmonized hippocampal segmentation protocol. Alzheimer's & Dementia **11**(2), 139–150 (2015)
2. Ataloglou, D., Dimou, A., Zarpalas, D., Daras, P.: Fast and precise hippocampus segmentation through deep convolutional neural network ensembles and transfer learning. Neuroinformatics **17**(4), 563–582 (2019). https://doi.org/10.1007/s12021-019-09417-y
3. Boccardi, M., et al.: Delphi definition of the EADC-ADNI harmonized protocol for hippocampal segmentation on magnetic resonance. Alzheimer's & Dementia **11**(2), 126–138 (2014)
4. Boccardi, M., et al.: Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. Alzheimer's & Dementia **11**(2), 175–183 (2015)
5. Breiki, F.A., Ridzuan, M., Grandhe, R.: Self-Supervised Learning for Fine-Grained Image Classification. arXiv preprint arXiv:2107.13973 (2021)
6. Carmo, D., Silva, B., Yasuda, C., Rittner, L., Lotufo, R.: Hippocampus segmentation on epilepsy and Alzheimer's disease studies with multiple convolutional neural networks. Heliyon **7**(2), e06226 (2021)
7. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations (2020). http://hdl.handle.net/20.500.11850/443425
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations (2020)
9. Du, A.T.: Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. J. Neurol. Neurosurg. Psychiat. **71**(4), 441–447 (2001)
10. Frisoni, G.B., et al.: The EADC-ADNI harmonized protocol for manual hippocampal segmentation on magnetic resonance: evidence of validity. Alzheimer's & Dementia **11**(2), 111–125 (2014)
11. Isensee, F., et al.: nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation (2018)
12. Jack, C.R., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magnet. Reson. Imaging **27**(4), 685–691 (2008)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012). https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
14. Mu, Y., Gage, F.H.: Adult hippocampal neurogenesis and its role in Alzheimer's disease. Molecul. Neurodegen. **6**(1), 85 (2011)
15. Navarro, F., et al.: Evaluating the Robustness of Self-Supervised Learning in Medical Imaging (2021)
16. Nobakht, S., Schaeffer, M., Forkert, N.D., Nestor, S., Black, S.E.P.B.: Combined Atlas and convolutional neural network-based segmentation of the hippocampus from MRI according to the ADNI harmonized protocol. Sensors **21**(7), 2427 (2021)
17. Noroozi, M., Favaro, P.: Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. arXiv preprint arXiv:1603.09246 (2016)

18. Petersen, R., et al.: Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology **74**(3), 201–209 (2010), cited By 913
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C.: QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage **186**, 713–727 (2019)
21. Setti, S.E., Hunsberger, H.C., Reed, M.N.: Alterations in hippocampal activity and Alzheimer's disease. Transl. Issue. Psychol. Sci. **3**(4), 348–356 (2017)
22. Sled, J., Zijdenbos, A., Evans, A.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging **17**(1), 87–97 (1998)
23. Smith, S.M.: Fast robust automated brain extraction. Human Brain Mapping **17**(3), 143–155 (2002)
24. Taleb, A., Lippert, C., Klein, T., Nabi, M.: Multimodal Self-Supervised Learning for Medical Image Analysis. arXiv preprint arxiv:1912.05396 (2019)