2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761627

SS-3DCAPSNET: SELF-SUPERVISED 3D CAPSULE NETWORKS FOR MEDICAL SEGMENTATION ON LESS LABELED DATA

Minh Tran^{*,1} *Loi Ly*† *Binh-Son Hua* ‡ *Ngan Le**

* University of Arkansas † Cyberlogitec Vietnam ‡ VinAI

ABSTRACT

Capsule network is a recent new deep network architecture that has been applied successfully for medical image segmentation tasks. This work extends capsule networks for volumetric medical image segmentation with self-supervised learning. To improve on the problem of weight initialization compared to previous capsule networks, we leverage self-supervised learning for capsule networks pre-training, where our pretext-task is optimized by self-reconstruction. Our capsule network, *SS-3DCapsNet*, has a UNet-based architecture with a 3D Capsule encoder and 3D CNNs decoder. Our experiments on multiple datasets including iSeg-2017, Hippocampus, and Cardiac demonstrate that our 3D capsule network with self-supervised pre-training considerably outperforms previous capsule networks and 3D-UNets. Code is available at here. [1]

Index Terms— Capsule network, self-supervised learning, medical image segmentation, less labeled data

1. INTRODUCTION

Since the introduction of UNet [1, 2], UNet-based neural networks have achieved impressive performance in various modalities of medical image segmentation (MIS), e.g. brain tumor [3, 4, 5], infant brain [6, 7], liver tumor [8], optic disc [9], retina [10], lung [11], and cell [12], etc. Recently, capsule networks [13] have also been applied successfully for MIS [14, 15, 11]. Despite such, there remains a wide range of challenges: (1) Most methods are based on supervised learning, which is prone to many data problems like small-scale data, low-quality annotation, small objects, ambiguous boundaries, to name a few. These problems are not straightforward to overcome: labeling medical data is laborious and expensive, requiring an expert's domain knowledge. (2) Capsule networks for medical segmentation does not outperform CNNs yet, even though the performance gap gets significantly closer [11].

To address such limitations and inspired by the recent success of capsule networks, in this work, we develop SS-3DCapsNet, a self-supervised capsule network for volumetric MIS. Our SS-3DCapsNet is built upon a state-of-the-art 3D capsule network that leverages both 3D Capsule blocks and CNN blocks for encoder and decoder architecture, respectively, which accounts for temporal relations in volumetric slices in learning contextual visual representation. We introduce self-supervised learning (SSL) to our 3D capsule network, which results in a UNet-like architecture that contains three pathways, i.e., visual representation, encoder, and decoder. The first path consists of dilated convolutional layers, which were pre-trained by SSL techniques. The encoder path is built upon 3D Capsule blocks, whereas the decoder path is built upon 3D CNNs blocks. Compared to 2D-SegCaps [15], which is highly dependent on some random phenomena such as sampling order or weight initialization, our SS-3DCapsNet learns visual representation better as well as having a more robust weight initialization thanks to self-supervised learning. Compared to 3D-UCaps [11], we show that self-supervised learning results in additional gain in segmentation accuracy while keeping the same network complexity at test time.

Our contributions are: (1) An effective self-supervised 3D capsules network for volumetric image segmentation. Our network architecture inherits the merits from 3D Capsule block, 3D CNN blocks, and self-supervised learning for better visual representation learning; and (2) A suite of experiments with ablation studies that empirically demonstrates the effectiveness of self-supervised 3D capsules network for MIS.

2. RELATED WORKS

Medical Segmentation. Among various DL architectures [16, 17], an encoder-decoder like UNet [1] and its extension have achieved impressive performance among semantic segmentation approaches. Since the seminal work of UNet [2] for MIS, there have been numerous subsequent works in this task. As shown in a recent survey [18], MIS can be divided into two main DL groups: supervised learning and weakly supervised learning techniques.

The first group includes CNN-based supervised learning methods such as FCN [19], UNet [20], CC-3D-FCN[21], RLS [22], ACRes[4], DenseVoxNet [23], Flow-based[24], VoxResNet [25], 3D DR-UNet [26], Recurrent Level Set [22], AtrousNet [4], Offset Curves Loss [7, 10], Point-Unet [5] as notable methods. The second group includes weakly supervised learning methods such as transfer learning [27], domain adaptation [28], interactive segmentation [29]. To address the issue of data limitation for training, Generative Adversarial Network

[1] Correspondence: minht@uark.edu

(GAN) [30] has been incorporated into CNNs [31, 32, 33, 24]. Training with imperfect datasets with scarce annotations and weak annotations has also been considered recently [34].

Capsule Networks. Capsule networks [35] (CapsNet) is a new network architecture concept that strengthens feature learning by retaining more information at the aggregation layer for pose reasoning and learning the part-whole relationship, which makes it a potential solution for semantic segmentation and object detection tasks. In CapsNet, a capsule aims to represent an entity: capsule norm indicates the probability that entity is present and capsule direction indicates the configuration that entity is in. CapsNet is recently made practical [13] in a CNN that incorporates two layers of capsules with dynamic routing.

While most CapsNet has been proposed for image classification, SegCaps [14, 15] expanded CapsNet for object segmentation. This method functions by treating an MRI image as a collection of slices, each of which is then encoded and decoded by capsules to output the segmentation. However, SegCaps is mainly designed for 2D still images, and it performs poorly when being applied to volumetric data because of missing temporal information. 3D-UCaps [11] is a hybrid network architecture that utilizes both capsules and deconvolutions for feature learning and segmentation output, respectively, which shows that such combination can outperforms SegCaps design significantly in the segmentation task while retaining the merits of capsules. Our method further improves upon 3D-UCaps by integrating an efficient pre-training stage.

Self-supervised Learning. Self-supervised learning (SSL) is a technique for learning feature representation in a network without requiring a labeled dataset. A common workflow to apply SSL is to train the network in an unsupervised manner by learning with a pretext task in the pre-training stage, and then finetuning the pre-trained network on a target downstream task. In the case of MIS, the suitable pretext tasks can be considered in four categories: context-based, generation-based, free semantic label-based, and cross-modal-based. The first techniques utilize context features of images or videos such as context similarity [36], spatial structure [37], temporal structure [38]. The second techniques have been used in image generation [39] and video generation [40]. The third techniques aim to automatically generate semantic labels and applied into segmentation [41], contour detection [41]. The fourth techniques are applied to multiple modalities data such as visual-audio [42], RGB-Flow [43]. In this work, our pretext task is based on image reconstruction.

3. SS-3DCAPSNET: SELF-SUPERVISED 3D CAPSULE NETWORKS

We draw on the ideas of SegCaps [14] and 3D-UCaps [11] to build our 3D capsule network for the medical segmentation task. Particularly, our network has three stages: (i) Visual representation, (ii) Capsule encoder, and (iii) Convolutional decoder as follows.

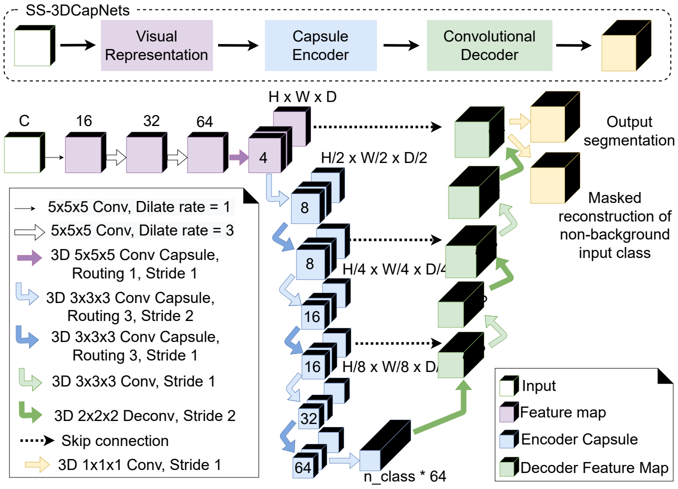

Fig. 1. Our proposed SS-3DCapsNet architecture with three components: visual representation; capsule encoder, and convolution decoder. Number on the blocks indicates number of channels in convolution layer and dimension of capsules in capsule layers.

(i) Visual Representation: This stage is for converting the input to a feature volume that can be consumed by the capsule encoder. Followed the concurrent work, we use three dilated convolution layers with 16, 32, 64 channels, respectively. The kernel size is set to $5 \times 5 \times 5$, with dilate rates of 1, 3, and 3, respectively. The size of the visual features is $H \times W \times D \times 64$.

(ii) Capsule Encoder: In this stage, we reshape the feature volume into $H \times W \times D$ capsules, where each capsule is represented by a 64-dimensional vector. Here we consider both spatial and temporal data by using our 3D convolutional capsules to learn a richer representation. The output from a convolution capsule has the shape $H \times W \times D \times C \times A$, where C is the number of capsule types and A is the dimension of each capsule. We follow the concurrent work and set C to $(16, 16, 16, 8, 8, 8)$ for each layer in the capsule encoder, respectively. Note that as the number of capsule types in the last convolutional capsule layer is equal to the number of class labels, we can further supervise this particular layer with a margin loss [13].

(iii) Convolutional Decoder: This is the final stage in our network. Here we use the decoder of 3D UNet [2] which includes deconvolution, skip connection, convolution and BatchNorm layers [44] to generate the segmentation from features learned by capsule layers. Particularly, we reshape the capsules back to tensors of size $H \times W \times D \times (C \star A)$ and pass them to the decoder. The overall architecture can be seen in Fig. 1.

3.1. Pretext Task

Our pretext task is self-supervised based on medical image reconstruction. In computer vision, it is common to use pseudo-

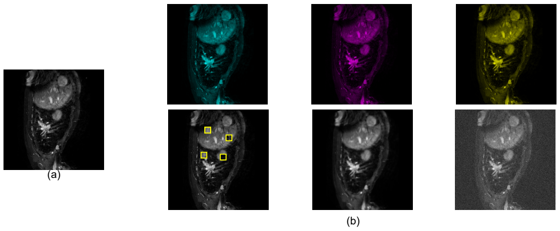

Fig. 2. Examples of six transformations for self-supervised learning. (a): original image. (b) from left to right, top to bottom: zeros-green-channel, zeros-red-channel, zeros-blue-channel, swapping (4 swapped patches are shown in yellow boxes), blurring, noisy.

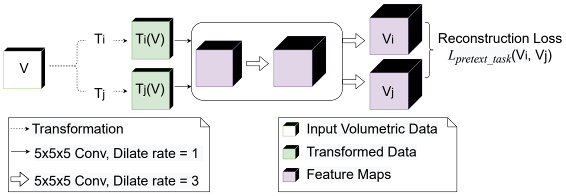

Fig. 3. Our pretext task with reconstruction loss.

labels defined by different image transformations, e.g, rotation, random crop, adding noise, blurring, scaling, flipping, jigsaw puzzle, etc. to supervise the pretext task. While such transformations work well for classification as a downstream task, since our downstream task is segmentation, we propose to use a pretext task that can consider reconstructing the original image. As medical images are captured in low contrast and the object-of-interest in medical images usually follows some specific patterns, we select contrast transformations to perform the pretext task with the reconstruction loss.

The details of our pre-training are as follows. Our pretext task is based on reconstruction from various transformations i.e. noisy, blurring, zero-channels (R,G,B), swapping as shown in Fig. 2. Let $\mathcal{F}$ is the visual representation network. The transformation is defined as $\{T_i\}_{i=1}^{i=N}$, where T_0 is an identity transformation and N is set as 6 corresponding to six transformations (Fig. 2). Let V denote as the original input volumetric data. Our pretext task is performed by applying two random transformations $T_i, T_j (i, j \in [0, 6])$ into V. The transformed data is then $T_i(V)$ and $T_j(V)$, respectively. The visual feature of transformed data after applying the network $\mathcal{F}$ is V_j and V_j, where $V_i = \mathcal{F}(T_i(V))$ and $V_j = \mathcal{F}(T_j(V))$. The network $\mathcal{F}$ is trained with a reconstruction loss defined by:

$$\mathcal{L}_{pretext}(V_i, V_j) = ||V_i - V_j||_2. \quad (1)$$

The pretext task procedure is illustrated in Fig. 3.

3.2. Downstream Task

After pre-training, we train our SS-3DCapsNet network with annotated data on the medical segmentation task. The total loss function to train this downstream task is a sum of three losses:

$$\mathcal{L}_{downstream} = \mathcal{L}_{margin} + \mathcal{L}_{CE} + \mathcal{L}_{reconstruction}. \quad (2)$$

The margin loss is adopted from [13] and it is defined between the predicted label y and the ground truth label y^* as follows:

$$\begin{aligned}\mathcal{L}_{margin} =& y^* \times (\max(0, 0.9 - y))^2 + \quad (3)\\ & 0.5 \times (1 - y^*) \times (\max(0, y - 0.1))^2.\end{aligned}$$

Particularly, we compute the margin loss ($\mathcal{L}_{margin}$) on the capsule encoder output with downsampled ground truth segmentation. We compute the weighted cross-entropy loss ($\mathcal{L}_{CE}$) on the convolutional decoder. We also regularize the training with a network branch that aims at reconstructing the original input with masked mean-squared errors ($\mathcal{L}_{reconstruction}$) [13, 14].

4. EXPERIMENTAL RESULTS

4.1. Implementation Details We conduct our experiments and comparisons on iSeg [45], Hippocampus, and Cardiac [46] datasets. For iSeg, we follow 3D-SkipDenseSeg [47] to have the training set of 9 subjects and testing set of subject #9. On Hippocampus, and Cardiac [46], the experiments are conducted by 4-fold cross-validation.

Table 1. Comparison on iSeg-2017. 1^{st} group: 3D CNN-based networks and 2^{nd} group: Capsule-based networks.

	Method	Depth	Dice Score			
			WM	GM	CSF	Average
CNN	Qamar et al. [48]	82	90.50	**92.05**	**95.80**	**92.77**
	3D-SkipDenseSeg [47]	47	**91.02**	91.64	94.88	92.51
	VoxResNet [25]	25	89.87	90.64	94.28	91.60
	3D-UNet [2]	18	89.83	90.55	94.39	91.59
	CC-3D-FCN [21]	34	89.19	90.74	92.40	90.79
	DenseVoxNet [23]	32	85.46	88.51	91.26	89.24
Capsule	2D SegCaps [14]	16	82.80	84.19	90.19	85.73
	3D-SegCaps [11]	16	86.49	88.53	93.62	89.55
	3D-UCaps [11]	17	90.21	91.12	**94.93**	92.08
	Our SS-3DCapsNet	17	**90.78**	**91.48**	94.92	**92.39**

We implemented our method in Pytorch. We used patch size of 64 × 64 × 64 for iSeg and Hippocampus whereas patch size of 128 × 128 × 128 on Cardiac. Our SS-3DCapsNet was trained without any data augmentation. We used Adam optimizer with an initial learning rate of 0.0001. The learning rate is decayed by 0.05 if the Dice score on the validation set does not increase for 50,000 iterations. Early stopping is set at 250,000 iterations as in [14].

Table 2. Comparison on Cardiac with 4-fold cross validation.

3D CNN-based networks		Capsule-based networks	
3D UNet[2]	84.30	SegCaps (2D) [14]	66.96
3D Vnet[49]	84.20	Multi-SegCaps (2D) [50]	66.96
3D DR-UNet [26]	87.40	3D-UCaps [11]	89.69
		Our SS-3DCapsNet	**89.77**

Table 3. Comparison on Hippocampus with 4-fold.

Method	Anterior			Posterior		
	Recall	Precision	Dice	Recall	Precision	Dice
Multi-SegCaps (2D) [50]	80.76	65.65	72.42	84.46	60.49	70.49
EM-SegCaps (2D) [50]	17.51	20.01	18.67	19.00	34.55	24.52
3D-UCaps [11]	81.70	80.19	80.99	80.2	79.25	79.48
Our SS-3DCapsNet	**81.84**	**81.49**	**81.59**	**80.71**	**80.21**	**79.97**

4.2. Performance and Comparison We compare our SS-3DCapsNet with both SOTA 3D CNNs-based and Capsule-based segmentation methods. 3D-Ucaps [11] has two versions of with and without utilizing MONAI [51]. To conduct a fair comparison, we report the version without MONAI.

The comparison between our proposed SS-3DCapsNet with SOTA segmentation approaches on iSeg dataset [45] is given in Table 1. As can be seen, 3D capsule networks (3D-SegCaps, SS-3DCapsNet) outperform 2D-SegCaps by a wide margin. This performance gap can be explained by the combination of pre-training, Capsule encoder, and Convolutional decoder in SS-3DCapsNet. Our SS-3DCapsNet also outperforms 3D-SegCaps, which contains only a Capsule-based encoder and decoder. Our SS-3DCapsNet also performs comparably to SOTA 3D CNNs, but our network is significantly shallower (17 layers vs. 82 layers in [48]). Our network also has fewer parameters and a better Dice score when compared to SOTA 3D CNNs with similar number of layers, e.g. 3D-UNets [2] (18 layers). In addition to iSeg, we also evaluate our SS-3DCapsNet on Hippocampus and Cardiac, where the results are shown in Table 2 and Table 3.

4.3. Ablation Study We analyze the performance of our method as follows.

i. Network Configuration: We trained SS-3DCapsNet under various settings as shown in Table 4. By following the concurrent work on 3D capsule networks, we use a baseline where the number of capsules of the first layer is reduced to 4 (similar to SegCaps). As can be seen, each component including visual representation, margin loss, reconstruction loss, pre-training contributes to the final performance. Removing any of such components would result in performance drops.

ii. SSL Contribution: We perform experiments on various datasets and turn on/off the self-supervision step in the experiments. The results in Table 5 clearly shows that pre-training plays an important role in our method, which improves the Dice score considerably in iSeg, and slightly in other datasets.

Table 4. Performance of SS-3DCapsNet on iSeg with different network configurations.

Method	Dice Score			
	WM	GM	CSF	Average
change number of capsule (set to 4)	89.02	89.78	89.95	89.58
w/o visual representation	89.15	89.66	90.82	89.88
w/o margin loss	87.62	88.85	92.06	89.51
w/o reconstruction loss	88.50	88.96	90.18	89.22
w/o pretext task	90.21	91.12	**94.93**	92.08
SS-3DCapsNet	**90.78**	**91.48**	94.92	**92.39**

Table 5. Performance of SS-3DCapsNet on Precision (Pre), Recall (Rec) and Dice score (DSC) with and without pretext task on various datasets.

	iSeg			Hippocamus			Cardiac		
	Pre	Rec	DSC	Pre	Rec	DSC	Pre	Rec	DSC
w/o. SSL	92.28	91.29	92.08	79.72	80.95	80.24	84.60	**95.06**	89.69
w. SSL	**92.54**	**92.37**	**92.39**	**80.85**	**81.27**	**80.78**	**86.24**	94.21	**89.77**

CONCLUSION

In this work, we proposed a capsule network for MIS powered with self-supervised pre-training. Our SS-3DCapsNet can both utilize self-supervised learning and 3D capsules for learning features while retaining the advantage of traditional convolutions in decoding the segmentation results. Even though we use capsules with dynamic routing only in the encoder of a simple Unet-like architecture, we can achieve the competitive result with the SOTA models on iSeg-2017 challenge while outperforming SegCaps [14] on different complex datasets with less labeled annotated data. Future work includes exploring different self-supervised learning methods such as SimCLR [52] for better feature learning and representation.

Acknowledgment This material is based upon work supported by the National Science Foundation under Award No. OIA-1946391 and NSF Track-2 WVAR-CRESH.

5. REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *MICCAI*, pp. 234–241, Springer, 2015.

[2] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3d u-net: learning dense volumetric segmentation from sparse annotation,” in *MICCAI*, pp. 424–432, Springer, 2016.

[3] L. Ngan, R. Gummadi, and M. Savvides, “Deep recurrent level set for segmenting brain tumors,” *MICCAI*, pp. 646–653, 2018.

[4] N. Le, K. Yamazaki, K. G. Quach, D. Truong, and M. Savvides, “A multi-task contextual atrous residual network for brain tumor detection & segmentation,” in *ICPR*, pp. 5943–5950, IEEE, 2021.

[5] N.-V. Ho, T. Nguyen, G.-H. Diep, N. Le, and B.-S. Hua, “Point-unet: A context-aware point-based neural network for volumetric segmentation,” in *MICCAI*, pp. 644–655, 2021.

[6] D.-H. Hoang, G.-H. Diep, M.-T. Tran, and N. T. Le, “Dam-al: Dilated attention mechanism with attention loss for 3d infant brain image segmentation,” *arXiv preprint arXiv:2112.13559*, 2021.

[7] N. Le, T. Le, K. Yamazaki, T. Bui, K. Luu, and M. Savides, “Offset curves loss for imbalanced problem in medical segmentation,” in *ICPR*, pp. 9189–9195, IEEE, 2021.

[8] P. Bilic *et al.*, "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056*, 2019.

[9] R. G. Ramani and J. J. Shanthamalar, "Improved image processing techniques for optic disc segmentation in retinal fundus images," *BSPC*, vol. 58, p. 101832, 2020.

[10] N. Le, T. Bui, V.-K. Vo-Ho, K. Yamazaki, and K. Luu, "Narrow band active contour attention model for medical segmentation," *Diagnostics*, vol. 11, no. 8, p. 1393, 2021.

[11] T. Nguyen, B.-S. Hua, and N. Le, "3d-ucaps: 3d capsules unet for volumetric image segmentation," in *MICCAI*, pp. 548–558, 2021.

[12] N. Moshkov, B. Mathe, A. Kertesz-Farkas, R. Hollandi, and P. Horvath, "Test-time augmentation for deep learning-based cell segmentation on microscopy images," *Scientific reports*, vol. 10, no. 1, pp. 1–7, 2020.

[13] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *arXiv preprint arXiv:1710.09829*, 2017.

[14] R. LaLonde and U. Bagci, "Capsules for object segmentation," *arXiv preprint arXiv:1804.04241*, 2018.

[15] R. LaLonde, Z. Xu, I. Irmakci, S. Jain, and U. Bagci, "Capsules for biomedical image segmentation," *MIA*, vol. 68, p. 101889, 2021.

[16] S. K. Zhou, H. N. Le, K. Luu, H. V Nguyen, and N. Ayache, "Deep reinforcement learning in medical imaging: A literature review," *Medical Image Analysis*, vol. 73, p. 102193, 2021.

[17] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, 2021.

[18] T. Lei, R. Wang, Y. Wan, X. Du, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *arXiv preprint arXiv:2009.13120*, 2020.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, pp. 3431–3440, 2015.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv e-prints*, p. arXiv:1505.04597, May 2015.

[21] D. Nie *et al.*, "3-d fully convolutional networks for multimodal isointense infant brain image segmentation," *IEEE trans on cybernetics*, vol. 49, no. 3, pp. 1123–1136, 2018.

[22] N. Le, R. Gummadi, and M. Savvides, "Deep recurrent level set for segmenting brain tumors," in *MICCAI*, pp. 646–653, Springer, 2018.

[23] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *CVPRW*, pp. 11–19, 2017.

[24] T. Bui, M. Nguyen, N. Le, and K. Luu, "Flow-based deformation guidance for unpaired multi-contrast mri image-to-image translation," in *MICCAI*, pp. 728–737, Springer, 2020.

[25] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images," *NeuroImage*, vol. 170, pp. 446–455, 2018.

[26] S. Vesal, N. Ravikumar, and A. Maier, "Dilated convolutions in neural networks for left atrial segmentation in 3d gadolinium enhanced-mri," in *STACOM-W*, pp. 319–328, Springer, 2018.

[27] A. A. Kalinin, V. I. Iglovikov, A. Rakhlin, and A. A. Shvets, "Medical image segmentation using deep neural networks with pre-trained encoders," in *Deep Learning Applications*, pp. 39–52, Springer, 2020.

[28] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," in *AAAI*, vol. 33, pp. 865–872, 2019.

[29] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren, "Deepigeos: a deep interactive geodesic framework for medical image segmentation," *IEEE TPAMI*, vol. 41, no. 7, pp. 1559–1572, 2018.

[30] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[31] Q. Chang, H. Qu, Y. Zhang, M. Sabuncu, C. Chen, T. Zhang, and D. N. Metaxas, "Synthetic learning: Learn from distributed asynchronized discriminator gan without sharing medical image data," in *CVPR*, pp. 13856–13866, 2020.

[32] N. Le, J. Sorensen, T. Bui, A. Choudhary, K. Luu, and H. Nguyen, "Enhance portable radiograph for fast and high accurate covid-19 monitoring," *Diagnostics*, vol. 11, no. 6, p. 1080, 2021.

[33] N. Le, J. Sorensen, T. D. Bui, A. Choudhary, K. Luu, and H. Nguyen, "Pairflow: Enhancing portable chest x-ray by flow-based deformation for covid-19 diagnosing," in *ICIP*, pp. 215–219, IEEE, 2021.

[34] N. Tajbakhsh *et al.*, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *MIA*, vol. 63, p. 101693, 2020.

[35] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *ICANN*, pp. 44–51, Springer, 2011.

[36] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *ECCV*, pp. 132–149, 2018.

[37] U. Ahsan, R. Madhok, and I. Essa, "Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition," in *WACV*, pp. 179–189, IEEE, 2019.

[38] D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman, "Learning and using the arrow of time," in *CVPR*, pp. 8052–8060, 2018.

[39] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*, pp. 649–666, Springer, 2016.

[40] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *ICML*, pp. 843–852, PMLR, 2015.

[41] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan, "Learning features by watching objects move," in *CVPR*, pp. 2701–2710, 2017.

[42] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *arXiv preprint arXiv:1807.00230*, 2018.

[43] N. Sayed, B. Brattoli, and B. Ommer, "Cross and learn: Cross-modal self-supervision," in *GCPR*, pp. 228–243, 2018.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, pp. 448–456, PMLR, 2015.

[45] L. Wang *et al.*, "Benchmark on automatic six-month-old infant brain segmentation algorithms: the iseg-2017 challenge," *IEEE TMI*, vol. 38, no. 9, pp. 2219–2230, 2019.

[46] A. L. Simpson *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

[47] T. D. Bui, J. Shin, and T. Moon, "Skip-connected 3d densenet for volumetric infant brain mri segmentation," *BSPC*, vol. 54, p. 101613, 2019.

[48] S. Qamar, H. Jin, R. Zheng, P. Ahmad, and M. Usama, "A variant form of 3d-unet for infant brain segmentation," *Future Generation Computer Systems*, vol. 108, pp. 613–623, 2020.

[49] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *CoRR*, 2016.

[50] S. Survarachakan, J. S. Johansen, M. Aarseth, M. A. Pedersen, and F. Lindseth, "Capsule nets for complex medical image segmentation tasks," *CVCS*, 2020.

[51] "Monai: Medical open network for ai." `https://monai.io`. Accessed: 2021-10-15.

[52] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.