A task-conditional mixture-of-experts model for missing modality segmentation

Philip Novosad[1,2], Richard A.D. Carano[1], and Anitha Priya Krishnan[1]

[1] Data Analytics and Imaging, Pharma Personalized Healthcare, Genentech Inc., South San Francisco, USA
[2] Hoffmann-La Roche Limited, Mississauga, Canada

Abstract. Accurate quantification of multiple sclerosis (MS) lesions using multi-contrast magnetic resonance imaging (MRI) plays a crucial role in disease assessment. While many methods for automatic MS lesion segmentation in MRI are available, these methods typically require a fixed set of MRI modalities as inputs. Such full multi-contrast inputs are not always acquired, limiting their utility in practice. To address this issue, a training strategy known as modality dropout (MD) has been widely adopted in the literature. However, models trained via MD still underperform compared to dedicated models trained for particular modality configurations. In this work, we hypothesize that the poor performance of MD is the result of an overly constrained multi-task optimization problem. To reduce harmful task interference, we propose to incorporate task-conditional mixture-of-expert layers into our segmentation model, allowing different tasks to leverage different parameters subsets. Second, we propose a novel online self-distillation loss to help regularize the model and to explicitly promote model invariance to input modality configuration. Compared to standard MD training, our method demonstrates improved results on a large proprietary clinical trial dataset as well as on a small publicly available dataset of T2 lesions.

Keywords: Multiple Sclerosis · Magnetic Resonance Imaging · Segmentation · Deep Learning · Missing modality · Multi-task learning · Mixture-of-experts · Self-distillation

1 Introduction

Multiple sclerosis (MS) is a chronic autoimmune disease characterized by demyelination and atrophy of the brain and spinal cord. Accurate quantification of T2 lesions through magnetic resonance imaging (MRI) plays a pivotal role in assessing disease activity and monitoring its progression. To comprehensively characterize MS lesions, a combination of complementary MR modalities such as T1w, T2w and fluid-attenuated inversion recovery (FLAIR) are often acquired.

While deep learning models have proven powerful for automated MS lesion segmentation, such models typically require a *fixed set* of MRI modalities as inputs, which can be impractical. For example, these models may be non-applicable in clinical settings where it is often the case that some modalities are missing due

to considerations of time/cost and/or image corruption/artifacts. A naive solution to this problem is to train dedicated models for each possible input modality configuration, but this too is impractical. A second disadvantage of fixed-input models is inefficient aggregation of training data across different datasets: either a limited set of common modalities must be selected (reducing available information) or a more complete set of modalities must be selected and all non-conforming data discarded (reducing the size of the training set).

It is therefore important to develop more generic segmentation models which can make optimal use of the available input configurations. While a number of different solutions have been proposed [1], the present work focuses on building upon a simple baseline training strategy known as *modality dropout* [14, 10]. Similar to conventional dropout, modality dropout randomly masks out modalities during training, forcing the model to avoid relying on interactions between modalities. Several extensions to modality dropout have been proposed, most notably through the use of specialized network architectures (e.g. multi-encoder single-decoder schemes [20, 5]) often accompanied by auxiliary tasks and loss functions (e.g. image reconstruction [20]). More similar to our work is a recent class of methods that explicitly use representations of the input modality configuration to drive feature fusion [5, 21] or dynamic feature interaction [12] schemes.

In the present work we describe a new method to improve the baseline modality dropout training strategy. We show that training with modality dropout can be seen as optimizing an overly constrained multi-task objective, where each unique modality configuration constitutes a task. To our knowledge, this is the first work to make an explicit connection between modality dropout and multi-task learning. To reduce harmful task interference, we propose the incorporation of task-conditional mixture-of-expert (MoE) [16] convolutional layers into our segmentation model, allowing different tasks (modality configurations) to leverage different model parameter subsets. Combined with a novel online self-distillation scheme, we demonstrate improved T2 lesion segmentation performance compared to standard modality dropout training, significantly reducing the performance gap compared to dedicated models trained specifically for particular modality configurations.

2 Methods

2.1 Modality dropout as multi-task learning

Consider a multi-modality image x with corresponding label y. The image x is represented as a tensor with K channels, i.e. $x = [x^1, ..., x^K]$. During training, modality dropout generates modality-dropped images $\tilde{x}$ on-the-fly by setting random tensor channels to zero:

$$\tilde{x}(x, m) = [x^1 m^1, ..., x^K m^K] \tag{1}$$

where $m \in \{0, 1\}^K$ is the corresponding modality code vector of length K describing the absence/presence of each of the K modalities. Training on modality-

dropped images $\tilde{x}$ discourages the model from relying on interactions between modalities, allowing a single shared model to handle all missing-modality conditions. To use a trained model for inference, given a query image, the input tensor is formed by placing the existing modalities into the appropriate channels, and substituting an equivalently sized array of zeros in place of missing modalities.

Let (x, y) be a sample from the training dataset D. Assuming the modality codes m are uniformly sampled during training from the set M containing all $2^K - 1$ unique modality codes, modality dropout minimizes the following loss function[3]:

$$\mathcal{L}_{md} = \mathbb{E}_{m\sim M}\left[\mathbb{E}_{(x,y)\sim D}\left[\mathcal{L}\Big(y, f_\theta(\tilde{x}(x,m))\Big)\right]\right] \tag{2}$$

where f_θ is the model with parameters θ, and $\mathcal{L}$ is a segmentation loss function.

Despite the simplicity of modality dropout, it often underperforms dedicated models trained for particular modality configurations. We argue that this behavior is expected if modality dropout is viewed through the lens of multi-task learning. Specifically from the formulation in equation (2), training with modality dropout can be recognized as optimizing a standard multi-task objective where each unique modality combination m constitutes a distinct task. In the multi-task optimization literature, it is known that conflicts arising from task differences can actually harm individual task performance, particularly when model parameters are extensively shared across tasks [19].

2.2 Task-conditional modeling with mixture-of-experts

To reduce interference between tasks, we propose a mixture-of-expert (MoE) modeling technique to enable task-conditional learning. Specifically, we replace a subset of convolutional layers in our network with MoE convolutional layers (Fig. 1). Each MoE layer consists of a set of N experts $\{E_1, ..., E_N\}$ and a gating network $\phi(m) : \{0, 1\}^K \rightarrow [0, 1]^N$ whose input is a binary modality code vector and whose output is an N-dimensional vector of normalized weights $\{w_1, ..., w_N\}$. The experts are standard convolutional layers of the same dimension, and the gating network is a linear layer operating on the modality code m followed by the softmax operation. The output of the MoE layer is given by a weighted summation of the outputs of the individual experts:

$$x_{out} = \sum_{n=1}^{N} \phi(m)_n E_n(x_{in}). \tag{3}$$

In this way, different experts can specialize on different tasks, reducing the extent of parameter sharing across tasks (see supplementary file for a visualization of trained expert weights). The formulation in equation (3) has the disadvantage of requiring N convolutions. As done in [18], we take advantage of the linearity

[3] If the modality codes are not uniformly sampled, the loss function in equation (2) need only be modified by the addition of task-specific weights which reflect relative sampling frequencies.

of convolution and instead mix the convolutional expert weights and biases first, and then apply a single convolution with the mixed parameters. Therefore, the MoE layer has the *same computational cost* as a standard convolutional layer.

Ideally, the experts in each MoE layer should specialize to different tasks, and each expert should be equally active on expectation over the different tasks. To explicitly encourage this behavior, we propose an additional regularization loss. Let $\phi^l(m)$ denote the vector of expert weights in response to modality code m at MoE layer l. The regularization loss for layer l is:

$$\mathcal{L}^l_{reg} = \mathrm{cv}(\mathbb{E}_{m\sim M}[\phi^l(m)]) + \mathbb{E}_{m\sim M}[\mathrm{cv}(\phi^l(m))] \tag{4}$$

where cv is the coefficient of variation. The first term encourages a uniform mean expert weight vector $\phi(m)$ across tasks, while the second term encourages expert specialization by promoting high variance among expert weights for a given task. The total regularization loss is given by the summation of the individual layer losses.

2.3 Online self-distillation

For convenience, we now consider the batch view, where we are given a batch of training images $\{(x_b, y_b)\}_{b=1}^B \sim D$. A corresponding batch of modality-dropped images $\{(\tilde{x}_b(x_b, m_b), y_b)\}_{b=1}^B$ is generated by randomly and uniformly sampling a modality code $m_b \sim M$ for each item in the batch. In our implementation, our segmentation loss term has contributions from both the original and modality-dropped images:

$$\mathcal{L}_{seg} = \frac{1}{B}\sum_{b=1}^{B}\mathcal{L}(y_b, f_\theta(x_b)) + \frac{1}{B}\sum_{b=1}^{B}\mathcal{L}(y_b, f_\theta(\tilde{x}_b)). \tag{5}$$

Inspired by recent self-distillation methods in self-supervised representation learning, we propose an additional loss function to regularize our models. Similar to [7, 4], this loss explicitly encourages the network to produce similar outputs given two views of the same input. In the context of our work, we treat the full-modality images x and the corresponding modality-dropped images $\tilde{x}$ as two such views. Given a batch $\{(x_b, y_b)\}_{b=1}^B$ and its corresponding modality-dropped version $\{(\tilde{x}_b(x_b, m_b), y_b)\}_{b=1}^B$, the self-distillation loss is:

$$\mathcal{L}_{sd} = \frac{1}{B}\sum_{b=1}^{B}\mathcal{L}_\tau(\mathrm{sg}(f_\theta(x_b)), f_\theta(\tilde{x}_b)) \tag{6}$$

where sg is the stop-gradient operator and the loss $\mathcal{L}_\tau$ signifies that the loss is computed on softened logits at a user-defined temperature τ. We note that while other works in missing modality segmentation have explored distilling representations from full-modality images [11, 6], our approach differs from these works by distilling knowledge on-the-fly within the *same model* throughout training,

without requiring a strong pre-trained teacher model. The total loss for our proposed method is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \alpha\mathcal{L}_{reg} + w(t)\beta\mathcal{L}_{sd} \tag{7}$$

where $w(t)$ is a sigmoid-like time-dependent ramp-up factor which takes on a maximum value of 1 after one-third of the total number of training steps [17].

3 Experiments

3.1 Datasets and Performance Measures

The **OPERA I** (NCT01247324) (https://clinicaltrials.gov/study/NCT01247324) clinical trial dataset is a large proprietary multi-site dataset containing T1w pre-contrast, T1w post-contrast, T2w, FLAIR scans sequences acquired axially with a resolution of $1 \times 1 \times 3$ mm^3. Ground truth T2 lesions were originally obtained by manual correction on the output of a semi-automated segmentation approach based on a Bayesian classifier. Because the original classifier used a multimodal input of T1w pre-contrast, T2w and FLAIR images, only those patients (n = 682) containing all such sequences and the corresponding annotations were included in this study. All images were previously pre-processed using non-uniformity correction and intra-patient rigid co-registration. The **MSSEG-2016** lesion challenge dataset [8] is a smaller (n = 53) multi-site publicly available dataset which includes T1w pre-contrast, T2w and FLAIR modalities. The ground truth T2 lesions were obtained by consensus among annotations from multiple experts. The dataset was previously pre-processed by the challenge organizers, with details are provided in [8]. We emphasize that no additional pre-processing was performed on either dataset by the authors of this work.

In addition to the Dice similiarity coefficient (DSC), we use a previously-validated composite score (CScore) [2] to measure the performance of the models. The CScore is defined as (DSC/8 + PPV/8 + LTPR/4 + (1-LFPR)/4 + Corr/8) where PPV is the positive predictive value, LTPR the lesion-wise true positive rate, LFPR the lesion-wise false positive rate, and Corr the volume correlation.

3.2 Implementation

We use a baseline 3D UNet [15] architecture for all experiments. Architectural details and the incorporation of MoE layers are shown in Fig. 1. All MoE layers use $n = 4$ experts. We study the impact of MoE layer placement in Section 3.3.

The MONAI [3] package is used for data-loading and inference. During training, all images are resampled to isotropic $1 \times 1 \times 1$ mm resolution and we sample random crops of dimension $96 \times 96 \times 96$ voxels with an 80% probability of center-cropping on a positive (non-background) pixel. Cropped patches are z-normalized independently per channel. For inference, a sliding-window technique is used to obtain the whole-image segmentation, after which the image is resampled back to native resolution.

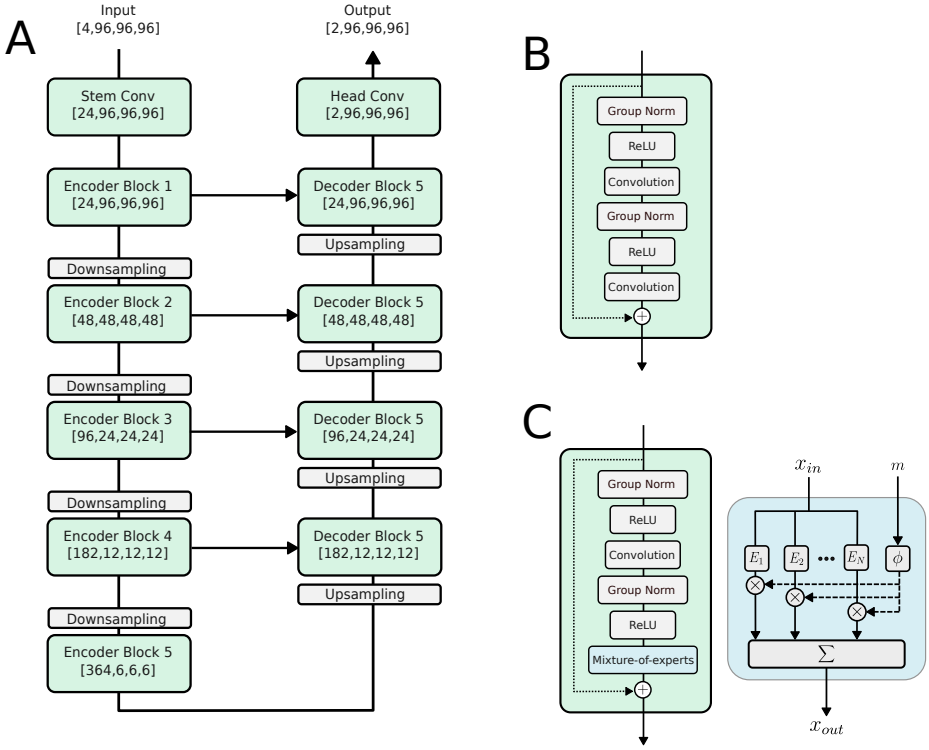

Fig. 1. Illustration of baseline 3D UNet model and incorporation of proposed MoE layers. **A**: The baseline model consists of an encoding path consisting of interspersed down-sampling (stride 2 convolution) and encoder blocks, and a decoding path, consisting of interspersed up-sampling blocks (stride 2 transposed convolution) and decoder blocks. The size of the output tensor corresponding to each block is shown (channel first format). All convolutions use 3×3×3 kernels. Both encoder and decoder blocks are pre-activation residual blocks (**B**). **C**: We adapt the residual blocks by replacing the second convolutional layer with a convolutional MoE layer, where the input/output channels of each convolutional layer expert E_n in the MoE layer are identical to those of the replaced convolutional layer.

All results were obtained by 3-fold cross-validation. Models were trained using PyTorch across multiple NVIDIA T4 GPUs using the Adam optimizer with initial learning rate $1e^{-4}$ with cosine decay and a batch size of 4. Owing to the order-of-magnitude difference between the size of the OPERA I (n = 682) and MSSEG-2016 (n = 53) datasets, we use 150k training steps for the former and 50k for the latter. We use an unweighted combined cross-entropy and soft Dice loss[4] as the segmentation loss $\mathcal{L}$ for both the supervised and self-distillation losses.

[4] The DiceCELoss class in the MONAI package was used. This exact formulation of the soft Dice loss can be found in [13].

Parameters α and β in equation (7) were set by default to 0.1 and 1 respectively. Self-distillation was carried out at a temperature $\tau = 2$. To help regularize the MoE layers and to encourage exploration across the various experts, following other works on MoEs [9, 16], we also add random noise with standard deviation 0.2 to the modality codes m during training. All above parameters were selected heuristically, but we empirically found that parameters τ and α were relatively insensitive in the ranges [1, 4] and [0.01, 1] respectively.

3.3 Results

We first study the contribution of and interactions between the components of the proposed method on the OPERA I dataset. Table 1 reports the performance of various model configurations. Replacing the UNet stem convolution with an MoE layer brings substantial gains, and further improvements can be obtained by incorporating MoE layers into each residual block (as shown in Fig. 1) in the encoding pathway of the UNet. Removing the regularization loss term results in a performance decrease. It is evident that each of the two principal contributions (incorporation of MoE layers and online self-distillation) bring substantial improvements individually, and that the best performance is obtained when both are used together. We also include a comparison with the work of [12] who propose a dynamic convolutional stem layer. Since the original implementation is for 2D convolutions, we re-implemented the dynamic layer in 3D and incorporated it into our baseline UNet model leaving all other variables unchanged. We confirm the findings of [12] that the dynamic layer improves performance over the baseline modality dropout model, but we find that the gains are relatively small.

Table 1. Impact of individual components of proposed model on the OPERA I dataset. Mean DSC and CScore with standard error across all $2^K - 1 = 7$ modality configurations are reported. The baseline modality dropout model is shown in the first row ($\mathcal{L}_{seg}$ only).

MoE in stem	MoE in encoder	$\mathcal{L}_{reg}$ ($\alpha = 0.1$)	$\mathcal{L}_{sd}$ ($\beta = 1$)	Dynamic stem [12]	DSC	CScore
-	-	-	-	-	66.94±0.21	70.80±0.17
-	-	-	-	✓	67.16±0.20	71.04±0.16
✓	-	✓	-	-	67.58±0.20	71.40±0.16
✓	✓	✓	-	-	67.66±0.20	71.58±0.16
✓	✓	-	-	-	67.39±0.20	71.40±0.16
-	-	-	✓	-	67.54±0.20	71.41±0.15
✓	✓	✓	✓	-	68.14±0.20	72.07±0.16

Table 2 reports the performance of three compared methods on the OPERA I dataset for various modality configurations. In general, the performance gap between standard modality dropout and dedicated models (models trained and

evaluated on fixed modality configurations) is reduced using our proposed method. When only a single modality is available, the improvements provided by our proposed method are larger, but there still remains a performance gap relative to dedicated models. When only a single modality is missing, the improvements provided by our proposed method are smaller, but the performance gap relative to the dedicated models is essentially closed. Aggregated over all modality configurations, Wilcoxon signed-rank tests confirm the improved performance of our method compared to standard modality dropout method ($p < 1e^{-8}$ for both DSC and CScore).

Table 2. Results on OPERA I dataset (n = 682). Mean DSC and CScore with standard error is reported. Full circles denote presence and empty circles absence. MD: modality dropout. DM: dedicated model trained specifically for a given modality configuration. For each metric and modality configuration, best and second best results are **bolded** and underlined respectively.

Modalities			DSC			CScore		
T1w	T2w	FLR	MD	Ours	DM	MD	Ours	DM
●	○	○	55.55±0.62	58.53±0.61	**60.42**±0.59	59.45±0.50	62.23±0.47	**63.91**±0.46
○	●	○	65.57±0.50	68.06±0.50	**69.57**±0.50	69.10±0.43	71.52±0.39	**72.82**±0.36
○	○	●	64.73±0.50	65.63±0.48	**66.29**±0.47	69.44±0.35	70.62±0.35	**70.88**±0.35
●	●	○	69.68±0.51	70.57±0.48	**71.13**±0.48	72.40±0.41	73.42±0.37	**73.87**±0.38
○	●	●	71.85±0.44	**72.35**±0.43	72.33±0.44	75.73±0.33	**76.53**±0.32	76.25±0.32
●	○	●	68.36±0.47	69.01±0.46	**69.09**±0.46	72.89±0.35	**73.44**±0.33	73.26±0.34
●	●	●	**72.80**±0.43	72.79±0.43	72.77±0.43	76.57±0.32	**76.75**±0.31	76.48±0.33

Finally, we reproduce our key findings using the MSSEG-2016 lesion dataset (n = 53). Table 3 reports the performance of the same three compared methods for various modality configurations. Aggregated over all modality configurations, Wilcoxon signed-rank tests confirm the improved performance of our method compared to standard modality dropout method ($p \leq 5e^{-8}$ for both DSC and CScore). Qualitative results are available in the supplementary file.

4 Discussion and Conclusion

In this work we demonstrate that treating modality dropout as a multi-task optimization problem can shed light on the sub-optimal performance of standard modality dropout, and also suggest interesting directions for improvement. Specifically, we show that using task-conditional MoE layers is an effective and general way to improve upon the modality dropout training strategy. However, the design space of MoE modeling (e.g. formulation of regularization loss, placement of MoE layers in residual block) is large and could not be fully explored in this short paper. Further investigation into the design of MoE layers and their

Table 3. Results on MSSEG-2016 dataset (n = 53). Mean DSC and CScore with standard error is reported. Full circles denote presence and empty circles absence. MD: modality dropout. DM: dedicated model trained specifically for a given modality configuration. For each metric and modality configuration, best and second best results are **bolded** and underlined respectively.

Modalities			DSC			CScore		
T1w	T2w	FLR	MD	Ours	DM	MD	Ours	DM
●	○	○	52.45±2.28	55.35±2.08	**57.13**±1.47	55.46±1.69	**58.24**±1.65	57.80±0.93
○	●	○	52.22±2.32	**55.28**±2.31	55.01±1.36	53.93±1.71	**57.37**±1.72	55.65±0.99
○	○	●	71.59±1.57	**72.28**±1.61	72.04±1.03	73.59±1.19	73.61±1.27	**74.14**±0.75
●	●	○	58.15±2.15	60.08±2.10	**61.77**±1.17	59.29±1.63	60.70±1.63	**60.77**±0.91
○	●	●	72.07±1.58	72.80±1.61	**72.87**±0.97	73.98±1.07	74.44±1.18	**74.45**±0.74
●	○	●	72.47±1.56	72.61±1.63	**72.78**±1.11	**74.61**±1.12	74.22±1.24	74.29±0.75
●	●	●	72.85±1.56	**72.97**±1.62	72.05±0.95	74.31±1.13	**74.57**±1.22	71.88±0.80

incorporation into different segmentation network architectures remains a topic for future research.

Acknowledgments. Special thanks to Alvaro Gomariz and Matthew McLeod for their valuable feedback.

Disclosure of Interests. This work was funded by Hoffmann-La Roche and Genentech Inc. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azad, R., Khosravi, N., Dehghanmanshadi, M., Cohen-Adad, J., Merhof, D.: Medical image segmentation on mri images with missing modalities: A review. arXiv preprint arXiv:2203.06217 (2022)
2. Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H.: Longitudinal multiple sclerosis lesion segmentation: resource and challenge. NeuroImage **148**, 77–102 (2017)
3. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
4. 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)
5. Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.A.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 447–456. Springer (2019)

6. Chen, C., Dou, Q., Jin, Y., Liu, Q., Heng, P.A.: Learning with privileged multimodal knowledge for unimodal segmentation. IEEE transactions on medical imaging **41**(3), 621–632 (2021)
7. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
8. Commowick, O., Kain, M., Casey, R., Ameli, R., Ferré, J.C., Kerbrat, A., Tourdias, T., Cervenansky, F., Camarasu-Pop, S., Glatard, T.: Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset. Neuroimage **244**, 118589 (2021)
9. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. The Journal of Machine Learning Research **23**(1), 5232–5270 (2022)
10. Feng, Y., Pan, H., Meyer, C., Feng, X.: A self-adaptive network for multiple sclerosis lesion segmentation from multi-contrast MRI with various imaging sequences. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 472–475. IEEE (2019)
11. Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., Gori, P.: Knowledge distillation from multi-modal to mono-modal segmentation networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 772–781. Springer (2020)
12. Liu, H., Fan, Y., Li, H., Wang, J., Hu, D., Cui, C., Lee, H.H., Zhang, H., Oguz, I.: Moddrop++: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 444–453. Springer (2022)
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
14. Neverova, N., Wolf, C., Taylor, G., Nebout, F.: Moddrop: adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8), 1692–1706 (2015)
15. 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)
16. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
17. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
18. Yang, B., Bender, G., Le, Q.V., Ngiam, J.: Condconv: Conditionally parameterized convolutions for efficient inference. Advances in neural information processing systems **32** (2019)
19. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. Advances in Neural Information Processing Systems **33**, 5824–5836 (2020)

20. Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 107–117. Springer (2022)
21. Zhao, Z., Yang, H., Sun, J.: Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 183–192. Springer (2022)