Self-supervised Learning of Radiomic Representations with a 3D Contrastive Masked Autoencoder

No Author Given

No Institute Given

Abstract. Self-supervised radiomics has attracted increasing attention as it learns data-driven representation complementary to traditional rule-based image features. Masked autoencoder (MAE) is a recent self-supervised representation learning technique based on vision transformer and shows promising results in the natural image domain. In this paper, we demonstrate their use in a long-standing problem in medical image computing: learning task-specific radiomic representations. To this end, first, we extend the *MAE* from 2D to 3D, which allows us to extract 3D representation from volumetric data. Second, we introduce an auxiliary task considering the structured dependencies to improve the reconstruction results and enhance the latent representation. Third, we introduce a new contrastive loss to strengthen the target representation directly. We evaluate the proposed method on two public multi-modal MRI datasets to quantify the characteristics of brain tumors. When combining the learned self-supervised feature with traditional radiomics, we demonstrate significant improvement in two diverse brain tumor classification tasks.

1 Introduction

Radiomics aims to offer a structured representation of medical imaging data for a given diagnostic task. Traditional radiomic features, defined by statistics of shape, texture and others [1], are proven to improve diagnosis in various tasks in the last decade. However, they are low-level (or rule-based) features and are not explicitly optimized on specific datasets, which limits their informativeness in representation. *Supervised* learning is optimized on specific tasks [9], but it is resource-consuming and often infeasible to gather large-scale annotated datasets in diverse tasks and image domains. As a consequence, it often overfits when datasets are limited and fails in many prediction tasks in practice [19].

Self-supervised representation learning (SSRL), especially contrastive learning-based methods [4,12,20], is a promising technique that can learn informative data representations *without* supervision. It is particularly useful and relevant in medical imaging, where expert knowledge is expensive and labeled data are often scarce for traditional supervised training. One popular way of performing *SSRL* is to attract *similar* image pairs and rebuff *dissimilar* image pairs. Given one batch, similar image pairs can be two augmented views of one image, and

the remaining samples can be used to construct the dissimilar image pairs. In medical imaging, *SSRL* are commonly used for two aspects: 1) pre-training deep networks for reduce labeling effort in the source domain [26] or downstream tasks [3,25], 2) extracting meaningful information from data, such as radiomics [15,8] which is the downstream task for applying *SSRL* in this work.

Most existing *SSRL* methods use *CNN*-based architecture (e.g. *ResNet*) as a backbone. As opposed to *CNN*, the recent Vision Transformer (ViT)-based approach called *masked autoencoder* (*MAE*) [11] is different with the above methods in principle. It randomly masks the sequential image patches and learns to reconstruct the original image (similar to image inpainting) using an autoencoder and *ViT* based architecture. The authors in [11] demonstrate that such a simple reconstruction can facilitate the Transformer to learn effective image representations in a self-supervised fashion. Given the promising results on natural images, our work focuses on adapting the 2D *MAE* to 3D imaging data and introducing new components to enhance the existing framework.

In the original *MAE* pipeline, it only computes pixel-wise reconstruction loss of the autoencoder. It does not consider structured dependencies of the predicted output, which might better capture the semantic features. For example, the structure of brain tumor in MR images can be hyper-intensities with rich texture and morphology, thus training with sole pixel-wise loss (often a $\mathcal{L}$-2 norm loss) might be insufficient for structural prediction. Another potential room for improving *MAE* is that - since the representation is learned by reconstruction via an autoencoder, it does not *directly* optimize the target representation. In contrast, contrastive learning-based methods [5,4] straightforwardly match the target representation from augmented views. This work aims to extend *MAE* and unite the two paradigms of *SSRL* methods to gain the complementary benefits of both approaches.

Contributions. Our contribution is threefold: (1) We introduce an auxiliary reconstruction task that considers the structural dependencies to enhance *MAE*. (2) For the first time, we combine Transformer- and contrastive learning-based methods to obtain an improved representation. (3) We demonstrate that self-supervised representations from multi-modal MR images can complement traditional radiomics for brain tumor patients in two diverse classification tasks.

2 Methodology

Overview. The objective is to learn a good domain-specific representation of volumetric data using a masked autoencoder without supervision. Consider an autoenconder with encoder E, decoder D, and a masking function $m(\cdot)$. Given an image volume X, we first decompose X into k sequential image patches with a size of $n \times \mathrm{n} \times n$. Then a random mask is applied on the sequential input patches to mask away $p\%$ of the patches. The remaining visible patches are referred to as X^*. The visible patches X^* are fed to the encoder F for learning features. For each missing patch, a token with 3D positional encoding is assigned to indicate the presence of such patches (called *MASK* tokens). The features of the

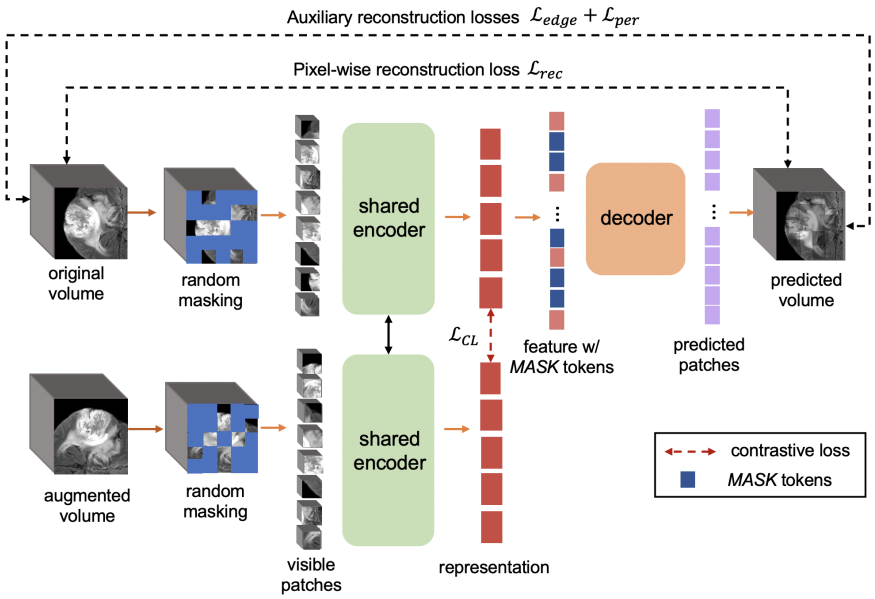

Fig. 1: Our proposed Contrastive Masked Autoencoder framework. The upper half of the figure shows the training procedure of masked autoencoder. A multi-modal MRI volume is masked randomly. The visible patches are passed through to the encoder. Placeholder tokens for the masked patches are introduced after the encoder (referred to as the MASK tokens). The full set of encoded features and these placeholder tokens are passed to a decoder which reconstructs the whole MRI volume. We use a pixel-wise reconstruction loss and two auxiliary losses based on edge map (Eq. 1) and perceptual features (Eq. 2). The lower half demonstrates the training with a contrastive loss. The input volume is augmented and masked randomly. Then it is passed through the shared encoder. The representations from two views are matched with cosine similarity loss (Eq. 3).

visible patches along with *MASK* tokens are passed to the decoder. The decoder G predicts image intensities for these *MASK* tokens and thereby reconstructs the whole volume, i.e., $X \approx D(E(X^*))$. We introduce an auxiliary edge-based reconstruction task, together with a loss function designed to capture high-order properties of the input to complement the pixel-wise loss. To enhance the target representation, we use contrastive loss. Fig. 1 shows the schematic view of our architecture. In the following sections, we explain each component in detail.

Masked autoencoder with pixel-wise loss. *MAE* reconstructs the original input given partial observations. A random masking strategy is used to mask $p\%$ of the input volume. Notably p is a hyperparameter which will be discussed in later section. The visible patches X^* are fed through encoder E. The decoder D reconstructs X from $E(X^*)$ using a mean-squared-error loss: $\mathcal{L}_{rec} = ||X - D(E(X^*)||_2$. Intuitively, the effectiveness of the representation should correlate with the reconstruction quality. Hence, two auxiliary losses are used to enhance the reconstruction.

Auxiliary reconstruction losses. Original *MAE* uses the pixel-wise loss for training as mentioned above. In our application of MRI images, the edges and perceptual features encode valuable supervision signals [21]. To provide a greater focus on this supervision signal during the training, we employ an edge-based loss and a perceptual loss based on high-level features. A 3D *Sobel* filter function [14] $Sobel(\cdot)$ is used to compute the gradients from three directions (see details in Appendix). The edge-based loss between original volume X and the reconstruction $\hat{X}$ can be formulated as:

$$\mathcal{L}_{edge} = ||Sobel(X) - Sobel(\hat{X})||_2 \tag{1}$$

To exact high-level features, we introduce VGG-based perceptual loss to compute the feature similarity at multiple levels over n 2D slices of one volume [13].

$$\mathcal{L}_{per} = \sum_{i=0}^{n} ||VGG(x_i) - VGG(\hat{x}_i)||_2 \tag{2}$$

where $VGG(\cdot)$ denotes multi-level features from the pre-trained VGG network.

Contrastive loss. Along with the reconstruction, a contrastive loss is introduced to enhance the target representation further. It tries to match the feature representations between the two views of visible patches, denoted as f_1 and f_2. We used negative cosine distance as the loss function and the optimization strategy in [4].

$$\mathcal{L}_{CL} = -\frac{f_1}{\|f_1\|_2} \cdot \frac{f_2}{\|f_2\|_2}, \tag{3}$$

Final objective loss. The final optimization objective is

$$L = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{edge} + \lambda_2 \mathcal{L}_{per} + \mathcal{L}_{CL} \tag{4}$$

where λ_1 and λ_2 are hyperparameters and initially set to 0.01 and 10 respectively considering the loss scale. λ_1 is decreased linearly to stabilize the training process. The motivation for the linear scaling and other hyperparameters are discussed in the Results section.

3 Experiments

Datasets and image preprocessing. The evaluation of our approach is performed on two public datasets: 1) a multi-center MRI dataset (*BraTS*) [18,2] including 326 patients with brain tumor. 2) The Erasmus Glioma Database (*EGD*) [24] which consists of 768 MRI scans. Both of the datasets include FLAIR, T1, T2 and T1-c with a uniform voxel size $1\times1\times1$ mm^3. We train individual models and evaluate on the two datasets separately. The effectiveness of the learned representations is evaluated on two down-stream classification tasks: a) For *BraTS*, discriminating high grade and low grade tumor, and b) for *EGD*, predicting isocitrate dehydrogenase (IDH) mutation status (0 or 1). For *EGD*, there are 768

scans in total, but only 468 scans have the information of IDH mutation status. Hence we use the 300 unlabeled scans for pre-training. Once trained, the encoder E is fixed and used to extract features for the 468 annotated scans.

We use the segmentation masks for both datasets to get its centroid and generate a 3D bounding box of $96 \times 96 \times 96$ to localize the tumor. If the bounding box exceeds the original volume, the out-of-box region is padded with background intensity. Four modalities are concatenated to server as a multi-modal input. The intensity range of all image volumes was rescaled to [0, 255] to guarantee the success of intensity-based transformations.

Architecture and training configuration. In this sub-section, we explain each component in the architecture and the training settings in detail.

Encoder and decoder. Working on only the visible tokens allows the encoder to process a fraction of 3D volumes and uses less GPU memory. We extend a standard ViT [7] to be applicable for 3D data and use it as the encoder. 3D patches are embedded using a linear projection with added positional encoding with 3D coordinates. 12 encoder blocks, each with an embedding dimension of 768, are used. The decoder reconstructs the volume using features from the encoder and the *MASK* tokens mentioned in Sec. 2. Each *MASK* token is a shared learned vector that indicates the presence of a missing patch to be predicted. We add 3D positional encoding to all the tokens for sequential prediction [6]. The decoder is realized using 8 transformer blocks, each with an embedding dimension of 512. After training, we only use the encoder as a feature extractor.

End-to-end training of MAE with contrastive loss. The MAE is trained for 1000 epochs using *AdamW* optimizer [17] with 0.05 weight decay. The base learning rate is 1e-3. The learning rate is annealed using cosine decay [16]. The batch size is set to 4, adjusted to maximize GPU memory with an Nvidia RTX 3090. We use 40 epochs for our warm-up schedule [10]. *Gamma correction, affine transform and Gaussian noise* augmentations are used to generate the second "view" of the input volume. The two views are randomly masked and passed through the shared encoder E. Features generated by the encoder are compared using their cosine similarity (Eq. 3).

Feature extraction. For each multi-modal volume, we extract a set of 107 traditional radiomics features [23] including first- and second-order statistics, shape-based features and gray level co-occurrence matrix, denoted as f_T. For the self-supervised learning one, we extract 768 features from the last transformer head of the encoder E, denoted as f_{ViT}. To directly evaluate the effectiveness of SSL-based features, we concatenate them to a new feature vector $f = [f_T, f_{ViT}]$. Note that f_T and f_{ViT} are always from the same subjects.

Evaluation strategy, classifier and metrics. For evaluation, we follow the standard strategy to evaluate the quality of the pre-trained representations by training a *supervised* linear support vector machine classifier on the training set and then evaluating it on the test set. We use the sensitivity, specificity

and Area Under the Receiver Operating Characteristic Curve (AUC) as the evaluation metrics. We use *stratified five-fold nested cross-validation* to reduce selection bias and validate each model. In each fold, we randomly sample 80% subjects from each class as the training set and the remaining 20% for each class as the test set. Furthermore, 20% of training data is separated and used exclusively to optimize the hyper-parameters within each fold.

4 Results

Quantitative comparison. We evaluate the effectiveness of the proposed self-supervised radiomics features on two classification tasks: a) discrimination of low grade and high grade of brain tumor and b) prediction of IDH mutation status.

Effectiveness of the auxiliary losses in reconstruction. On *BraTS* dataset, the features learned by vanilla self-supervised method behave rather poorly compared to traditional radiomics (0.331 *vs.* 0.628 in sensitivity). The introduced auxiliary losses (edge loss plus perceptual loss) improves the representation, i.e., 0.382 *vs.* 0.331 in sensitivity and 0.734 *vs.* 0.696 in AUC. In Fig. 2, one could observe that the proposed auxiliary losses improve the reconstruction results, which indicates that the latent representation carries more semantic information compared to using only L2 norm loss. Notably, although the reconstruction is imperfect, it improves the performance on the downstream applications.

Complementary to existing radiomics features. From the first row of Tab. 1, one can observe that traditional radiomics itself brings discriminativeness to quantify tumor characteristics. The aggregation of the vanilla self-supervised representation and traditional radiomics shows a significant improvement in specificity (from 0.747 to 0.844) and AUC. Importantly, after including the auxiliary reconstruction losses and contrastive learning, we observe an increase in both specificity and sensitivity of the model with a higher AUC (rows 8, 9 & 10). Compared to traditional radiomics alone, the proposed solution with auxiliary reconstruction losses and contrastive loss shows an increase in specificity by **8.9%** (83.6% *vs.* 74.7%), while sensitivity is increased by **13.3%** (76.1% *vs.* 62.8%) and AUC by **6.4%** (0.876 *vs.* 0.812) on the *BraTS* dataset. A similar improvement is observed for the *EGD* dataset, where the specificity increases by 7.1 % (88.8% *vs.* 81.7%), sensitivity by 1.3% (80.6% *vs.* 79.3%) and AUC by 1.2% (0.922 *vs.* 0.910). Our approach also outperforms a recent contrastive learning-based method (3D SimSiam) [15] (row 2 in Tab. 1).

Comparison with supervised methods. To compare our proposed solution with a supervised model, we train a 3D ResNet [22] on the input volumes. Data augmentations used for contrastive learning and class weighting are employed to train the network properly. The 3D ResNet exhibits good specificity (82.2% *vs.* 74.7% from traditional radiomics features) but struggles with lower sensitivity and lower AUC due to limited labeled training data. Our method outperforms the supervised one on all three metric on both datasets.

Analysis of hyperparameters. We analyze two critical parameters in the proposed framework which affect the representation and training stability.

Table 1: Comparison of performances of different features in two down-stream tasks using stratified cross-validation. Our proposed self-supervised representation complements traditional radiomicss and outperforms supervised learning. (Spec.=Specificity, Sens.=Sensitivity, blue cells indicate the best two among all methods)

	BraTS [18]			*EGD* [24]		
Methods	*Spec.*	*Sens.*	*AUC*	*Spec.*	*Sens.*	*AUC*
Trad. radiomics	0.747	0.628	0.812	0.817	0.793	0.910
3D SimSiam [15]	0.812	0.667	0.839	0.826	0.784	0.912
MAE	0.812	0.331	0.696	0.887	0.600	0.838
MAE+Edge	0.820	0.318	0.704	0.875	0.651	0.846
MAE+Edge+Per.	0.812	0.382	0.734	0.884	0.607	0.814
MAE+Edge+Per.+CL	0.804	0.422	0.767	0.887	0.529	0.822
Trad. + MAE	**0.844**	0.604	0.827	0.817	0.813	0.899
Trad.+MAE+Edge	0.812	0.709	0.822	0.826	**0.817**	0.918
Trad.+MAE+Edge+Per.	0.836	0.667	0.853	0.887	0.784	0.918
Trad.+MAE+Edge+Per.+CL	0.836	**0.761**	**0.876**	**0.888**	0.806	**0.922**
Supervised learning	0.822	0.458	0.785	0.859	0.671	0.867

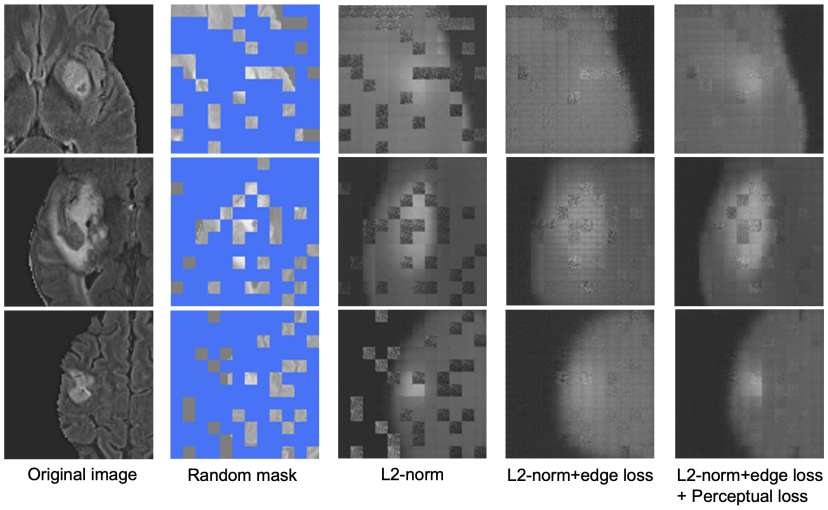

Fig. 2: Effect of different loss functions for image reconstruction. We observe the progressive improvement of the reconstructed results from a fraction (25%) of the input.

Effect of masking ratio p. In general 3D image data are natural signals with high spatial redundancy. The masking ratio p determines what percentage of the input volume is masked away. This, in turn, controls the difficulty of the reconstruction task. A higher value of p implies fewer visible patches X^*, which makes reconstruction more challenging. On the other hand, a low p can allow the model to extrapolate between visible patches, thus not learning good features. To find suitable value of p, we run experiments with $p \in [0.55, 0.65, 0.75, 0.85, 0.95]$. The results for the *BraTS* dataset are summarized in Figure 3. We observe that

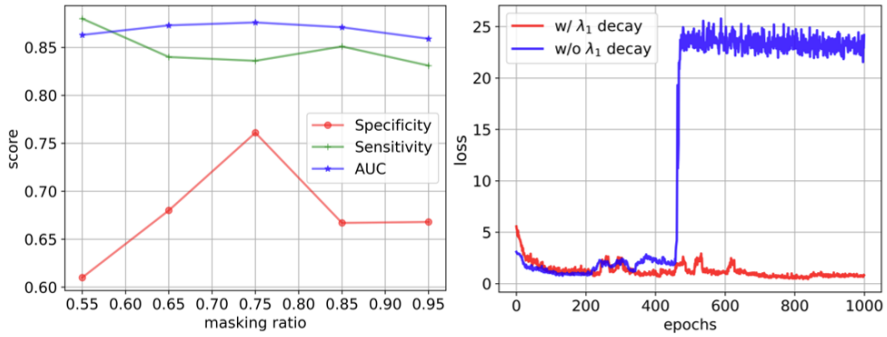

Fig. 3: Effects of two key parameters. ***Left:*** *Masking ratio* vs. *specificity, sensitivity and AUC on the BraTS dataset. The AUC and specificity increase with higher masking ratio, peaking at 0.75, after which they begin to decrease. Generally, the sensitivity slightly decreases with higher ratio.* ***Right:*** *Training loss with Eq. 4 over epochs with and without weight decay of λ_1 for $\mathcal{L}_{edge}$. Without decay (blue), the loss may explode and the training could not converge; With decay (red), training is stable and converges.*

the highest AUC value is obtained at $p = 0.75$. In our experiments, the optimal value was $p = 0.75$ for both datasets.

Weight decay for edge-based loss. The edge-based loss weight λ_1 in Eq. 4 was decreased linearly with number of epochs. We observed that gradually decreasing λ_1 was essential for the training stability. One possible explanation for the instability is related to the one-to-multiple mapping of the reconstruction task. While the edge map $\mathcal{L}_{edge}$ provides strong additional supervision signals for reconstruction, it is sensitive to noise artifacts which can cause large loss. Fitting to such artifacts destabilizes the training process and hampers the embedding features. To avoid this overfitting, we start the training with $\lambda_1 = 0.01$, and linearly decrease it with each epoch. This allows the network to learn structural information initially and gradually focus more on the perceptual similarity. The loss plot is shown in Fig. 3. Without a weight decay of λ_1, the loss tends to diverge. When using a linear decay, the training is stable and gradually converges.

5 Conclusion

To overcome existing limitations of radiomic features, we proposed a self-supervised learning approach to extract quantitative information. We introduce a Vision Transformer-based *SSRL* method for characterizing volumetric data. For the first time, we extend the existing *MAE* framework by introducing two auxiliary reconstruction losses and combining it with a contrastive learning-based method. We demonstrate that our proposed method enjoys the benefits of both worlds. The new self-supervised radiomics can complement traditional radiomics. When pre-training only on unlabelled data, our method outperforms supervised learning-based methods. We hope that our work can provide a new perspective for representation learning in medical imaging and advance radiomic features to the next level.

References

1. Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications **5**(1), 1–9 (2014)
2. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
3. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. arXiv preprint arXiv:2006.10511 (2020)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
5. 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)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
8. Dufumier, B., Gori, P., Victor, J., Grigis, A., Wessa, M., Brambilla, P., Favre, P., Polosan, M., Mcdonald, C., Piguet, C.M., et al.: Contrastive learning with continuous proxy meta-data for 3d mri classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 58–68. Springer (2021)
9. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639), 115–118 (2017)
10. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour (2018)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
12. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
14. Kanopoulos, N., Vasanthavada, N., Baker, R.L.: Design of an image edge detection filter using the sobel operator. IEEE Journal of solid-state circuits **23**(2), 358–367 (1988)
15. Li, H., Xue, F.F., Chaitanya, K., Luo, S., Ezhov, I., Wiestler, B., Zhang, J., Menze, B.: Imbalance-aware self-supervised learning for 3d radiomic representations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 36–46. Springer (2021)

16. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts (2017)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019)
18. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2014)
19. Mutasa, S., Sun, S., Ha, R.: Understanding artificial intelligence based radiology studies: What is overfitting? Clinical imaging **65**, 96–99 (2020)
20. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
21. Shen, H., Wang, R., Zhang, J., McKenna, S.J.: Boundary-aware fully convolutional network for brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 433–441. Springer (2017)
22. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition (2018)
23. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer research **77**(21), e104–e107 (2017)
24. van der Voort, S.R., Incekara, F., Wijnenga, M.M., Kapsas, G., Gahrmann, R., Schouten, J.W., Dubbink, H.J., Vincent, A.J., van den Bent, M.J., French, P.J., et al.: The erasmus glioma database (egd): Structural mri scans, who 2016 subtypes, and segmentations of 774 patients with glioma. Data in brief **37**, 107191 (2021)
25. Zeng, D., Wu, Y., Hu, X., Xu, X., Yuan, H., Huang, M., Zhuang, J., Hu, J., Shi, Y.: Positional contrastive learning for volumetric medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 221–230. Springer (2021)
26. Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J.: Models genesis: Generic autodidactic models for 3d medical image analysis. In: Medical Image Computing and Computer-Assisted Intervention. pp. 384–393 (2019)