Multimodal Self-supervised Learning for Medical Image Analysis

Aiham Taleb[1](✉), Christoph Lippert[1], Tassilo Klein[2], and Moin Nabi[2]

[1] Hasso-Plattner Institute, Potsdam University, Potsdam, Germany
{aiham.taleb,christoph.lippert}@hpi.de
[2] SAP AI Research, Berlin, Germany
{tassilo.klein,m.nabi}@sap.com

Abstract. Self-supervised learning approaches leverage unlabeled samples to acquire generic knowledge about different concepts, hence allowing for annotation-efficient downstream task learning. In this paper, we propose a novel self-supervised method that leverages multiple imaging modalities. We introduce the *multimodal* puzzle task, which facilitates representation learning from multiple image modalities. The learned modality-agnostic representations are obtained by confusing image modalities at the data-level. Together with the Sinkhorn operator, with which we formulate the puzzle solving optimization as permutation matrix inference instead of classification, they allow for efficient solving of multimodal puzzles with varying levels of complexity. In addition, we also propose to utilize generation techniques for multimodal data augmentation used for *self-supervised pretraining*, instead of downstream tasks directly. This aims to circumvent quality issues associated with synthetic images, while improving data-efficiency and the representations learned by self-supervised methods. Our experimental results show that solving our multimodal puzzles yields better semantic representations, compared to treating each modality independently. Our results also highlight the benefits of exploiting synthetic images for self-supervised pretraining. We showcase our approach on three segmentation tasks, and we outperform many solutions and our results are competitive to state-of-the-art.

Keywords: Self supervised learning · Multimodal images analysis

1 Introduction

Modern medical diagnostics heavily rely on the analysis of multiple imaging modalities, e.g. for differential diagnosis. However, to leverage the data for supervised machine learning approaches, it requires annotation of large numbers of training examples. Generating expert annotations of patient multimodal data at scale is non-trivial, expensive, time-consuming, and is associated with risks on privacy leakages. Consequently, scarcity of data annotations is one of the main impediments for machine learning applications in medical imaging. Self-supervised learning provides a viable solution when labeled training data is

A. Feragen et al. (Eds.): IPMI 2021, LNCS 12729, pp. 661–673, 2021.
https://doi.org/10.1007/978-3-030-78191-0_51

scarce. In these approaches, the supervisory signals are derived from the data itself, typically by the unsupervised learning of a proxy task. Subsequently, the obtained models facilitate data-efficient supervised fine-tuning on target real-world downstream tasks, hence reducing the burden of manual annotation.

Many self-supervised methods utilize the spatial context as a rich supervisory signal to learn effective data representations. However, these approaches neglect an important characteristic of medical images: their multimodality, e.g. MRI and CT. From an anatomical perspective, multimodality is essential because differences in the physical properties of organs and tissues are translated in a complementary fashion across multiple modalities. Examples of such cases are numerous [9]: soft body tissues are better encoded in MRI, but CT scans capture bone structures better. Likewise, specific brain tissues or tumors are better seen in specific MRI modalities, and so on. Thus, we propose to include multiple imaging modalities in our multimodal Jigsaw puzzle task, to integrate the cross-modal information in learned representations using a modality confusion loss.

While the learned representations by our multimodal puzzles prove useful in several downstream tasks when trained and fine-tuned using *realistic* multimodal images, as shown in Sects. 4.2 and 4.3. We also propose to utilize a cross-modal generation step to enhance the quantities of multimodal samples used in training the puzzle solver. Not only this step allows for better adoption of our multimodal puzzles in real-world scenarios, but also demonstrates the possibility of utilizing *synthetic* images for self-supervised *pretraining*, instead of downstream task training. This step is motivated by clinical scenarios, where data is often non-registered, and the quantities of modalities may vary, i.e. creating a modality imbalance problem. By introducing this step in our pipeline, the imbalance issue is alleviated, as shown in Sect. 4.4.

Our Contributions are two-fold. First, a novel self-supervised multimodal puzzle-solving task, which mixes multiple imaging modalities at the data-level, allowing for combining the cross-modal complementary information about different concepts in the data. Second, we propose to exploit cross-modal image generation (translation) for self-supervised tasks, instead of downstream tasks directly. Our results show that exploiting inexpensive solutions similar to ours can provide gains in medical image analysis, particularly in low data regimes.

2 Related Work

Self-supervised learning methods differ in the supervision source used to create the proxy tasks. Common supervision sources include the spatial context [8,18], image colors [28], clustering [5], image rotation prediction [11], image reconstruction [29], and contrastive learning [7,19]. We propose a novel spatial context derived across multiple imaging modalities, encouraging the model to learn modality-agnostic notions from the data. Noroozi *et al.* [18] proposed to solve Jigsaw puzzles on natural images. In contrast to our approach, their method relies on a single imaging modality, limiting its ability to capture vital cross-modal information. In addition, their method requires massive memory and compute resources, even for small puzzles. We improve the computational tractability by employing the Sinkhorn operator [21] as an analog to the Softmax in permutation tasks, allowing to solve more complex puzzles. This operator casts puzzle solving as a permutation matrix inference instead of classification. In their method, the choice of a fixed permutation set as classification targets limits the self-supervised task complexity. Instead, by defining our task as a matrix inference, the model searches among *all* permutations.

Solving jigsaw puzzles can be employed in a multi-task fashion, e.g. [4, 22], as a secondary task for domain adaptation. In this scenario, the model is expected to confuse the modalities/domains at the *feature-level*, similar to the late modality fusion [3]. In contrast, we fuse the modalities in our generated puzzles, i.e. performing a *data-level* early fusion. As opposed to our approach, their approach is likely to fail when the modality difference is high, as shown in our experiments.

In the medical context, self-supervision has been applied to depth estimation in medical image registration [16], body part recognition [27], and in disc degeneration using spinal MRIs [14]. These works make assumptions about input data, resulting in engineered solutions that hardly generalize to other target tasks. Our proposed approach avoids such assumptions about the data. Instead, our results show that it may operate on different imaging modalities, even when spatially unregistered. In more related works, Tajbakhsh *et al.* [23] use orientation prediction from medical images as a proxy task, Zhou *et al.* [29] employ image reconstruction techniques, and Taleb *et al.* [24] extend several self-supervised methods to 3D medical scans. Zhuang *et al.* [31] develop a proxy task for solving 3D jigsaw puzzles, in an attempt to simulate Rubik's cube solving. Since their work extends the 2D puzzles of Noroozi *et al.* [18] to 3D, it incurs similar computational costs, only the issue is exacerbated here as 3D puzzles require more computations. We follow this line of research, and we exploit multiple imaging modalities and improve puzzle solving efficiency in our proxy task.

Image-to-image translation using Generative Adversarial Networks [13, 30] has found several use cases in the medical domain. Many works have attempted to translate across medical imaging modalities [25,26]. While the goal of these methods is to improve the cross-modal generation quality, we view it as orthogonal to our goal. Similar to [10], we utilize cross-modal translation methods to improve the performance on downstream tasks, e.g. segmentation, through augmentation. However, especially for clinical applications, one can doubt the quality of synthetic images. Hence, as opposed to the above works, we circumvent this issue by employing these images for pretraining purposes only.

3 Method

Our method processes multimodal images, as is the case in many medical imaging datasets [9]. We assume no prior knowledge about what modalities are being used in our models, i.e. the modalities can vary from one task to another.

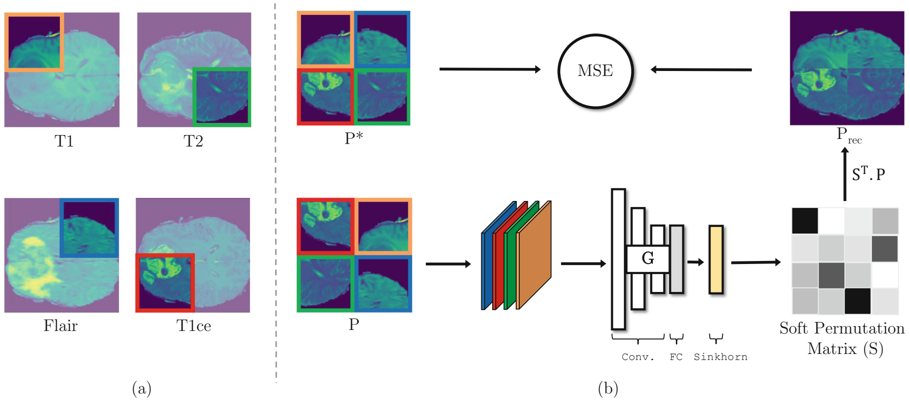

Fig. 1. Schematic illustration showing the proposed multimodal puzzles. (a) Assuming we have four modalities (this number can vary), (b) these images are then used to construct multimodal jigsaw puzzles, drawing patches from the modalities randomly.

3.1 Multimodal Puzzle Construction

Solving a jigsaw puzzle entails two steps. First, the image is cut into puzzle pieces (patches or tiles), which are shuffled according to a certain permutation. Second, these shuffled pieces are assembled such that the original image is restored. If N is the number of puzzle pieces, then there exist $N!$ of possible arrangements. In a conventional puzzle, the tiles originate from one image at a time, i.e. the computational complexity is $O(N!)$. On the other hand, we propose a *multimodal* jigsaw puzzle, which simultaneously learns the in-depth representation of how organs compose, along with the cross-modal spatial relationships. Hence, the tiles can stem from M different modalities. As a result, the computational complexity is increased to $O(N!^M)$. This quickly becomes prohibitively expensive due to: i) factorial growth in the number of permutations $N!$, ii) exponential growth in the number of modalities M. To reduce the computational burden, we use two tricks. First, we employ the Sinkhorn operator, which efficiently addresses the factorial growth. Second, we employ a feed-forward network G that learns a rich cross-modal representation, which cancels out the exponential factor M.

3.2 Puzzle-Solving with Sinkhorn Networks

To efficiently solve our multimodal jigsaw puzzle task, we train a network that can learn a permutation. A permutation matrix of size $N \times N$ corresponds to one permutation of the numbers 1 to N. Every row and column, therefore, contains precisely a single 1 with 0s everywhere else, which is a non-differentiable parameterization. However, it can be approximated in terms of a differentiable relaxation, the so-called Sinkhorn operator [21]. The Sinkhorn operator iteratively normalizes rows and columns of any real-valued matrix to obtain a "soft" permutation matrix, which is doubly stochastic. Formally, for an arbitrary N

dimensional square matrix X, the Sinkhorn operator $S(X)$ is defined as:

$$\begin{aligned} S^0(X) &= exp(X), \\ S^i(X) &= \mathcal{T}_R(\mathcal{T}_C(S^{i-1}(X))), \\ S(X) &= \lim_{i \to \infty} S^i(X). \end{aligned} \tag{1}$$

where $\mathcal{T}_R(X) = X \oslash (X \mathbf{1}_N \mathbf{1}_N^\top)$ and $\mathcal{T}_C(X) = X \oslash (\mathbf{1}_N \mathbf{1}_N^\top X)$ are the row and column normalization operators, respectively. The element-wise division is denoted by $\oslash$, and $\mathbf{1}_N^\top \in \mathbb{N}^N$ is an N dimensional vector of ones.

Assuming an input set of patches $P = \{p_1, p_2, ..., p_N\}$, where $P \in \mathbb{R}^{N \times l \times l}$ represents a puzzle that consists of N square patches, and l is the patch length. We process each patch in P with a network G, which produces a single output feature vector with length N per patch. By concatenating together these feature vectors obtained for all region sets, we obtain an $N \times N$ matrix, which the Sinkhorn operator converts to the soft permutation matrix $S \in [0, 1]^{N \times N}$. Formally, the network G learns the mapping $G : P \to S$, by minimizing the mean squared error (MSE) between the sorted ground-truth patch set P^* and the reconstructed version P_{rec} of the scrambled input. Then, S is applied to the scrambled input P to reconstruct the image P_{rec}, as in the formula:

$$\mathcal{L}_{puzzle}(\theta, P, P^*) = \sum_{i=1}^{K} \left\| P_i^* - S_{\theta, P_i}^T . P_i \right\|^2, \tag{2}$$

where θ are the network parameters, and K is the total number of training puzzles. The network parameters in G encode the cross-modal representations of different tissue structures. Therefore, they can be employed in downstream tasks by fine-tuning them on target domains in an annotation-efficient regime. Our approach is depicted in Fig. 1.

3.3 Cross-modal Generation

Multimodal medical scans exist in several curated datasets, and in pairs of aligned (or registered) scans. However, in many real-world scenarios, obtaining such data in large quantities can be challenging. To address this, we add an explicit cross-modal generation step using CycleGAN [30]. This model also uses a cycle-consistency loss, which relaxes the alignment (pairing) constraint across the two modalities, requiring no prior expensive registration. This step allows for leveraging the richness of multimodal representations obtained by our proposed puzzle-solving task. In our experiments, we generate samples of the smaller modality (in number of samples) using the larger modality. Then, we construct our multimodal puzzles using a mix of real and generated multimodal data. As we show in our related experiments in Sect. 4.4, this yields better representations compared to using a single modality only when creating puzzles.

4 Experimental Results

In Sect. 4.1, we provide details about the datasets. In Sect. 4.2, we evaluate the learned representations by assessing their impact on downstream tasks performance. Also, we compare to baselines from state-of-the-art. In Sect. 4.3, we assess the obtained benefits in data-efficiency. Finally, in Sect. 4.4, we analyze how data generation affects the learned representations. The experiments in the Sects. 4.3 and 4.4 are performed in ablation mode.

4.1 Datasets

We consider three multimodal medical imaging datasets. The first is the Brain Tumor Image Segmentation Benchmark (**BraTS**) [1,17], which contains multimodal MRI scans for 285 training and 66 validation cases. All cases include four MRI modalities: T1, T1Gd, T2, and T2-FLAIR volumes. The second is for **Prostate** segmentation [20], which consists of 48 multimodal MRI cases (32 for training, and 16 for testing). Segmentation masks of the prostate were produced from T2 scans and ADC maps. The third is for **Liver** segmentation from the CHAOS [15] dataset, which consists of 40 multimodal cases (20 for training, and 20 for testing). This dataset contains CT and MRI (T2) modalities, where each case has one scan per modality. The modalities in this benchmark are also non-registered, making it a pertinent test-bed for our multimodal puzzles.

4.2 Transfer Learning Results[1]

Brain Tumor Segmentation. This task involves segmenting 3 regions of brain tumor: a) whole tumor (WT), b) tumor core (TC), and c) enhanced tumor (ET).

Baselines: apart from the `Single-modal` baseline, all of the baselines use multimodal data for pretraining[2]: i) `From Scratch:` provides an insight into the benefits of pretraining, as opposed to learning the target task directly. ii) `Single-modal:` studies the impact of pretraining using a single modality as input. We employ the best modality for each task, i.e. Flair for whole tumor, T2 for tumor core, and T1ce for enhanced tumor. iii) `Isensee` ***`et al.`*** [12]: This model ranks among the tops in the BraTS 2018 challenge. `2D Isensee` is a 2D version of their network, for better comparability. iv) `Chang` ***`et al.`*** [6]: Trained multiple versions of 2D U-Nets and formed an ensemble, which requires significantly more compute time and resources than our single model. v) `JiGen` [4]: is a multi-tasking approach, which solves jigsaw puzzles as a secondary task. It aims to analyze the benefits of data-level modality confusion (ours), as opposed to feature-level (JiGen). vi) `Models Genesis` [29]: is a self-supervised method that uses image reconstruction. We compare to the 2D version (`2D MG`).

[1] We Evaluate on *realistic* data in this section, using a 5-fold cross validation approach.

[2] In fine-tuning, we use the same multimodal data across all models.

Evaluation Metrics. The reported metrics are the average dice scores for the Whole Tumor (WT), the Tumor Core (TC), and the Enhanced Tumor (ET).

Discussion. The results of our multi-modal method compared to the above baselines are shown in Table 1. Our method outperforms both the from scratch and single-modal baselines, confirming the benefits of pretraining using our multimodal approach. In addition, our method outperforms the baselines of Chang *et al.* [6] and 2D Isensee *et al.* [12], even though the latter uses co-training with additional datasets and augmentation techniques. This supports the benefits of initializing CNNs with our multimodal puzzles. We also outperform 2D Models Genesis (2D MG) [29] in this downstream task, supporting the effectiveness of our pretraining method. Compared to JiGen [4], we also find that our approach of performing the modality confusion in the data-level is superior to modality confusion in the feature-level, in this downstream task.

Table 1. Results on three segmentation benchmarks

Dataset	BraTS			Prostate				Liver
Model	Dice			Dice		NSD		Dice
	ET	WT	TC	C	P	C	P	
From scratch	68.12	80.54	77.29	68.98	86.15	94.57	97.84	89.98
Single-modal	78.26	87.01	82.52	69.48	87.42	92.97	97.21	92.01
Registered	–	–	–	–	–	–	–	95.09
Chang [6]	75.90	89.05	82.24	–	–	–	–	–
2D Isensee [12]	78.92	88.42	83.73	–	–	–	–	–
2D MG [29]	79.21	88.82	83.60	73.85	87.77	94.61	98.59	95.01
JiGen [4]	78.75	88.15	83.32	69.98	86.82	92.67	96.13	93.18
Ours (Multi-modal)	**79.65**	**89.74**	**84.48**	**75.11**	**88.79**	**94.95**	**98.65**	**95.10**

Prostate Segmentation. The target of this task is to segment two regions of the prostate: central gland and peripheral zone.

Baselines apart from Single-modal, all use multimodal data. We evaluate similarly to brain tumor segmentation, and we compare to the same baselines, only here we train on prostate data: i) From Scratch. ii) Single-modal: on MRI (T2). iii) JiGen [4]. iv) Models Genesis (2D MG) [29].

Evaluation Metrics. We compute the average Dice score and the normalized surface distance (NSD), for both prostate regions (**C**entral and **P**eripheral).

Discussion. The results of our multi-modal method compared to the above baselines are shown in Table 1. Our method outperforms both from scratch and single-modal baselines in this task, too, supporting the advantages of pretraining using our method. We also outperform 2D Models Genesis (2D MG) [29] in this task, supporting the effectiveness of our pretraining method. Our method outperforms JiGen [4] when trained on this task too. We observe a larger gap in performance between our approach and JiGen in this task, compared to brain tumor segmentation. We posit that this is due to the more significant difference between the modalities in this task, which can be seen clearly in Fig. 3, and is caused by the underlying physics. ADC maps measure water diffusion within organ tissues (the prostate exhibits lower diffusion). BraTS modalities are variants of T1- and T2-weighted MRI, which only differ in scanner configurations.

Liver Segmentation. Using unregistered abdominal CT and MRI scans.

Baselines apart from `Single-modal`, all use multimodal data. We evaluate similarly to brain tumor segmentation, and we compare to the same baselines, only here we train on liver data: i) `From Scratch`. ii) `Single-modal`: we employ CT to create the puzzles. iii) `JiGen` [4]. iv) `2D Models Genesis (2D MG)` [29]. v) `Registered`: to assess the influence of registration on our method, we register the modalities in this baseline (using VoxelMorph [2])[3].

Evaluation Metrics. We compute the average dice score (Dice) for liver.

Discussion. The results of our `multimodal` method compared to the above baselines are shown in Table 1. Our method outperforms both `from scratch` and `single-modal` baselines in this task too, supporting the advantages of pre-training using our method. We outperform 2D Models Genesis (`2D MG`) [29] only marginally in this task, and we believe this is because Models Genesis was pre-trained on Chest CT. Also, our method outperforms `JiGen` [4], when trained on this task too. The results against the `Registered` baseline are almost on par with the our method (trained on non-registered data). This result highlights our multimodal puzzles' generalization ability to non-registered modalities.

4.3 Low-Shot Learning Results

In this set of experiments, we assess how our self-supervised task benefits data-efficiency, by measuring the performance on downstream tasks at different labeling rates, i.e. fine-tuning pre-trained models on corresponding sample sizes. We randomly select subsets of patients at 1%, 10%, 50%, and 100% of the total training set size. For each subset, we compare the performance of our `multimodal` model to the baselines `from scratch` and `single-modal`. As shown in Fig. 2, our method outperforms both. The performance gap to the `single-modal` baseline confirms the benefits of using our *multimodal* puzzles. In a low-data regime of 1% of the dataset size, the margin to the `from scratch` baseline appears larger. This case, in particular, suggests the potential for generic unsupervised features applicable to medical imaging tasks, and has consequences on annotation efficiency. We report these results on *realistic* data. The `single-modal` baselines use: FLAIR for BraTS, T2 for Prostate, and CT for Liver.

[3] Our aim is to benchmark our method against a proven image registration method.

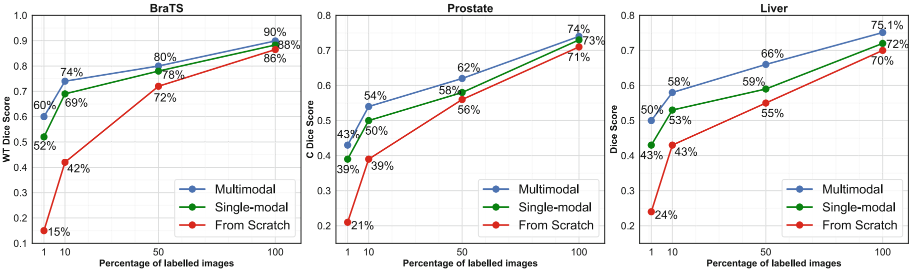

Fig. 2. Results in the low-shot scenario. Our method improves data-efficiency.

4.4 Cross-modal Generation Results

We study in these experiments the effect of the cross-modal generation step, which is motivated by the imbalance in imaging modality quantities. Hence, in this set of experiments, we perform this step in a semi-supervised fashion, assuming small multimodal and large single-modal subsets. We evaluate at multimodal subset sizes of 1%, 10%, and 50% of the total number of patients in each benchmark. We assume a reference modality[4], which is often abundant in practice, to generate the other modalities. In BraTS, since we have four MRI modalities, we train three GANs and convert T2-MRI to T1-, T1CE-, and FLAIR-MRI. In Prostate, we use T2-weighted MRI scans to generate the ADC diffusion-weighted scans. In CHAOS, we use the CT modality to generate T2 MRI.

Discussion. This step is justified if it provides a performance boost over the `single-modal` baseline, i.e. training on puzzles from one modality. The presented results of `Our method` in Table 2 clearly show an improvement on all benchmarks, when training our puzzle solver on a mixture of synthetic and realistic multimodal data. Even when we use only 1% of dataset sizes, the generator appears to capture the important characteristics of the generated modality. The qualitative results in Fig. 3 confirm the quality of generated images. In addition, the results in Table 2 support the benefits of using the synthetic data for self-supervised pretraining, instead of `Downstream training` directly.

[4] Alternatively, all modalities can be generated from each other, requiring many GANs.

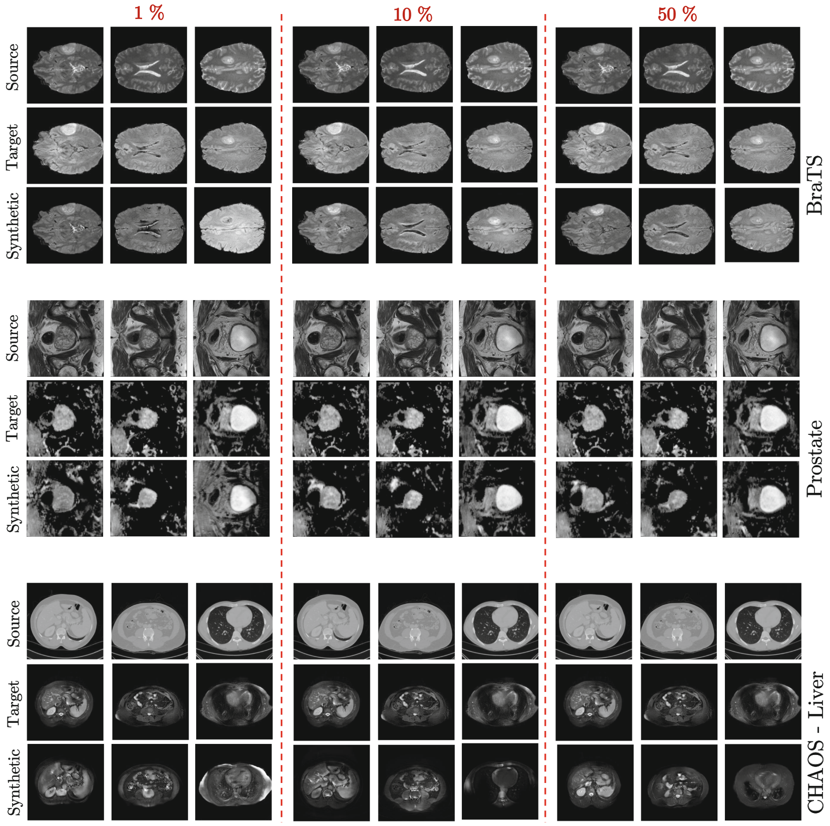

Fig. 3. Qualitative results of CycleGAN at different multimodal data rates, which affects generation quality. Here, translation is from T2 to FLAIR in BraTS, T2 to ADC in Prostate, and CT to T2-MRI in CHAOS (its targets are obtained with VoxelMorph)

Table 2. Segmentation results (in dice score). The rates are sizes of multimodal subsets

Model	BraTS			Prostate		CHAOS
	ET	WT	TC	C	P	Liver
Single-modal	72.12	82.72	79.61	69.48	87.42	92.01
Downstream training (1%)	65.40	74.11	69.24	55.24	71.23	80.31
Downstream training (10%)	69.28	78.72	71.58	62.65	76.18	83.65
Downstream training (50%)	72.92	81.20	78.36	66.34	80.24	87.58
Our method (1%)	73.12	82.42	80.01	61.87	82.67	82.71
Our method (10%)	74.19	85.71	81.33	67.67	84.37	86.26
Our method (50%)	**76.23**	**87.04**	**82.38**	**73.45**	**87.92**	**93.85**

5 Conclusion and Future Work

In this work, we proposed a self-supervised *multimodal* Jigsaw puzzle-solving task. This approach allows for learning rich semantic representations that facilitate downstream task solving in the medical imaging context. The proposed multimodal puzzles outperform their single-modal counterparts, confirming the advantages of including multiple modalities when constructing jigsaw puzzles. We also showed competitive results to state-of-the-art in three medical imaging benchmarks. One of which has unregistered modalities, further supporting the effectiveness of our approach in producing rich data representations. In addition, our approach further reduces the cost of manual annotation required for downstream tasks, and our results in the low-data regime support this benefit. We also evaluated a cross-modal translation method as part of our framework, which when used in conjunction with our method, it showed performance gains even when using few multimodal samples to train the generative model. Finally, we demonstrated the benefits of multimodality in our multimodal jigsaw puzzles, and we aim to generalize this idea to other self-supervised tasks. In addition, we believe generalizing our multimodal puzzles to the 3D context should improve the learned representations, and we deem this as a future work.

References

1. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 1–13 (2017)
2. Balakrishnan, G., Zhao, A., Sabuncu, M., Guttag, J., Dalca, A.: Voxelmorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging 1 (2019). https://doi.org/10.1109/TMI.2019.2897538
3. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (2019)
4. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: CVPR, IEEE (2019)
5. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: The European Conference on Computer Vision (ECCV). Springer, Munich, Germany (September 2018)
6. Chang, Y.J., Lin, Z.S., Yang, T.L., Huang, T.Y.: Automatic segmentation of brain tumor from 3d MR images using a 2d convolutional neural networks. In: Pre-Conference Proceedings of the 7th MICCAI BraTS Challenge, Springer (2018)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (2020)
8. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV, pp. 1422–1430. IEEE, USA (2015)
9. Eisenberg, R., Margulis, A.: A Patient's Guide to Medical Imaging. Oxford University Press, New York (2011)
10. Fu, C., et al.: Three dimensional fluorescence microscopy image synthesis and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (06 2018). https://doi.org/10.1109/CVPRW.2018.00298
11. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. CoRR arXiv abs/1803.07728 (2018)

12. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No new-net. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 234–244. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_21
13. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976. IEEE, Honolulu, Hawaii, USA (2017)
14. Jamaludin, A., Kadir, T., Zisserman, A.: Self-supervised learning for spinal MRIs. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 294–302. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_34
15. Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S.: CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data (April 2019). https://doi.org/10.5281/zenodo.3362844
16. Li, H., Fan, Y.: Non-rigid image registration using self-supervised fully convolutional networks without training data. In: ISBI, pp. 1075–1078. IEEE (April 2018)
17. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)
18. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
19. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. CoRR arXiv abs/1807.03748 (2018)
20. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. CoRR arXiv abs/1902.09063 (2019)
21. Sinkhorn, R.: A relationship between arbitrary positive matrices and doubly stochastic matrices. Ann. Math. Stat. **35**(2), 876–879 (1964)
22. Sun, Y., Tzeng, E., Darrell, T., Efros, A.A.: Unsupervised domain adaptation through self-supervision (2019)
23. Tajbakhsh, N., et al.: Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1251–1255 (2019)
24. Taleb, A., et al.: 3d self-supervised methods for medical imaging. In: NeurIPS (2020)
25. Wolterink, J.M., Dinkla, A.M., Savenije, M.H.F., Seevinck, P.R., van den Berg, C.A.T., Išgum, I.: Deep MR to CT synthesis using unpaired data. In: Tsaftaris, S.A., Gooya, A., Frangi, A.F., Prince, J.L. (eds.) SASHIMI 2017. LNCS, vol. 10557, pp. 14–23. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68127-6_2
26. Yang, H., et al.: Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 174–182. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_20
27. Zhang, P., Wang, F., Zheng, Y.: Self-supervised deep representation learning for fine-grained body part recognition. In: ISBI, pp. 578–582. IEEE (April 2017)
28. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 649–666. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_40
29. Zhou, Z., et al.: Models genesis: generic autodidactic models for 3D medical image analysis. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 384–393. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_42

30. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251. IEEE, Venice, Italy (2017)
31. Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., Zheng, Y.: Self-supervised feature learning for 3D medical images by playing a Rubik's cube. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 420–428. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_46