

# 3D Self-Supervised Methods for Medical Imaging

Aiham Taleb <sup>1,\*</sup>, Winfried Loetzsch <sup>1,†</sup>, Noel Danz <sup>1,†</sup>, Julius Severin <sup>1,\*</sup>, Thomas Gaertner <sup>1,†</sup>, Benjamin Bergner <sup>1,\*</sup>, and Christoph Lippert <sup>1,\*</sup>

<sup>1</sup>Digital Health & Machine Learning, Hasso-Plattner-Institute, Potsdam University, Germany

\*{firstname.lastname}@hpi.de

†{firstname.lastname}@student.hpi.uni-potsdam.de

## Abstract

Self-supervised learning methods have witnessed a recent surge of interest after proving successful in multiple application fields. In this work, we leverage these techniques, and we propose 3D versions for five different self-supervised methods, in the form of proxy tasks. Our methods facilitate neural network feature learning from *unlabeled* 3D images, aiming to reduce the required cost for expert annotation. The developed algorithms are 3D Contrastive Predictive Coding, 3D Rotation prediction, 3D Jigsaw puzzles, Relative 3D patch location, and 3D Exemplar networks. Our experiments show that pretraining models with our 3D tasks yields more powerful semantic representations, and enables solving downstream tasks more accurately and efficiently, compared to training the models from scratch and to pretraining them on 2D slices. We demonstrate the effectiveness of our methods on three downstream tasks from the medical imaging domain: i) Brain Tumor Segmentation from 3D MRI, ii) Pancreas Tumor Segmentation from 3D CT, and iii) Diabetic Retinopathy Detection from 2D Fundus images. In each task, we assess the gains in data-efficiency, performance, and speed of convergence. Interestingly, we also find gains when transferring the learned representations, by our methods, from a large unlabeled 3D corpus to a small downstream-specific dataset. We achieve results competitive to state-of-the-art solutions at a fraction of the computational expense. We publish our implementations<sup>1</sup> for the developed algorithms (both 3D and 2D versions) as an open-source library, in an effort to allow other researchers to apply and extend our methods on their datasets.

## 1 Introduction

Due to technological advancements in 3D sensing, the need for machine learning-based algorithms that perform analysis tasks on 3D imaging data has grown rapidly in the past few years [1–3]. 3D imaging has numerous applications, such as in Robotic navigation, in CAD imaging, in Geology, and in Medical Imaging. While we focus on medical imaging as a test-bed for our proposed 3D algorithms in this work, we ensure their applicability to other 3D domains. Medical imaging plays a vital role in patient healthcare, as it aids in disease prevention, early detection, diagnosis, and treatment. Yet efforts to utilize advancements in machine learning algorithms are often hampered by the sheer expense of the expert annotation required [4]. Generating expert annotations of 3D medical images at scale is non-trivial, expensive, and time-consuming. Another related challenge in medical imaging is the relatively small sample sizes. This becomes more obvious when studying a particular disease, for instance. Also, gaining access to large-scale datasets is often difficult due to privacy concerns. Hence, scarcity of data and annotations are some of the main constraints for machine learning applications in medical imaging.

<sup>1</sup><https://github.com/HealthML/self-supervised-3d-tasks>

Several efforts have attempted to address these challenges, as they are common to other application fields of deep learning. A widely used technique is transfer learning, which aims to reuse the features of already trained neural networks on different, but related, target tasks. A common example is adapting the features from networks trained on ImageNet, which can be reused for other visual tasks, e.g. semantic segmentation. To some extent, transfer learning has made it easier to solve tasks with limited number of samples. However, as mentioned before, the medical domain is supervision-starved. Despite attempts to leverage ImageNet [5] features in the medical context [6-9], the difference in the distributions of natural and medical images is significant, i.e. generalizing across these domains is questionable and can suffer from dataset bias [10]. Recent analysis [11] has also found that such transfer learning offers limited performance gains, relative to the computational costs it incurs. Consequently, it is necessary to find better solutions for the aforementioned challenges.

A viable alternative is to employ self-supervised (unsupervised) methods, which proved successful in multiple domains recently. In these approaches, the supervisory signals are derived from the data. In general, we withhold some part of the data, and train the network to predict it. This prediction task defines a proxy loss, which encourages the model to learn semantic representations about the concepts in the data. Subsequently, this facilitates data-efficient fine-tuning on supervised downstream tasks, reducing significantly the burden of manual annotation. Despite the surge of interest in the machine learning community in self-supervised methods, only little work has been done to adopt these methods in the medical imaging domain. We believe that self-supervised learning is directly applicable in the medical context, and can offer cheaper solutions for the challenges faced by conventional supervised methods. Unlabelled medical images carry valuable information about organ structures, and self-supervision enables the models to derive notions about these structures with no additional annotation cost.

A particular aspect of most medical images, which received little attention by previous self-supervised methods, is their 3D nature [12]. The common paradigm is to cast 3D imaging tasks in 2D, by extracting slices along an arbitrary axis, e.g. the axial dimension. However, such tasks can substantially benefit from the full 3D spatial context, thus capturing rich anatomical information. We believe that relying on the 2D context to derive data representations from 3D images, in general, is a suboptimal solution, which compromises the performance on downstream tasks.

**Our contributions.** As a result, in this work, we propose five self-supervised tasks that utilize the full 3D spatial context, aiming to better adopt self-supervision in 3D imaging. The proposed tasks are: 3D Contrastive Predictive Coding, 3D Rotation prediction, 3D Jigsaw puzzles, Relative 3D patch location, and 3D Exemplar networks. These algorithms are inspired by their successful 2D counterparts, and to the best of our knowledge, most of these methods have never been extended to the 3D context, let alone applied to the medical domain. Several computational and methodological challenges arise when designing self-supervised tasks in 3D, due to the increased data dimensionality, which we address in our methods to ensure their efficiency. We perform extensive experiments using four datasets in three different downstream tasks, and we show that our 3D tasks result in rich data representations that improve data-efficiency and performance on three different downstream tasks. Finally, we publish the implementations of our 3D tasks, and also of their 2D versions, in order to allow other researchers to evaluate these methods on other imaging datasets.

Contrastive Predictive Coding- <https://arxiv.org/abs/1807.03748v2>

Rotation prediction- <https://dl.acm.org/doi/10.1145/3348445.3348485>

Jigsaw- <https://arxiv.org/abs/1603.09246>

Patch Location- <https://arxiv.org/pdf/1505.05192>

## 2 Related work

In general, unsupervised representation learning can be formulated as learning an embedding space, in which data samples that are semantically similar are closer, and those that are different are far apart. The self-supervised family constructs such a representation space by creating a supervised proxy task from the data itself. Then, the embeddings that solve the proxy task will also be useful for other real-world downstream tasks. Several methods in this line of research have been developed recently, and they found applications in numerous fields [13]. In this work, we focus on methods that operate on images only.

Self-supervised methods differ in their core building block, i.e. the proxy task used to learn representations from unlabelled input data. A commonly used supervision source for proxy tasks is the spatial context from images, which was first inspired by the skip-gram Word2Vec [14] algorithm. This idea was generalized to images in [15], in which a visual representation is learned by predicting the position of an image patch relative to another. A similar work extended this patch-based approach to solve

**Jigsaw Puzzles** [16]. Other works have used different supervision sources, such as image colors [17], clustering [18], image rotation prediction [19], object saliency [20], and image reconstruction [21]. In recent works, **Contrastive Predictive Coding (CPC) approaches** [22, 23] advanced the results of self-supervised methods on multiple imaging benchmarks [24, 25]. These methods utilize the idea of contrastive learning in the latent space, similar to **Noise Contrastive Estimation** [26]. In 2D images, the model has to predict the latent representation for next (adjacent) image patches. Our work follows this line of research in the above works, however, our methods utilize the full 3D context.

While **videos** are rich with more types of supervisory signals [27–31], we discuss here a subset of these works that utilize 3D-CNNs to process input videos. In this context, 3D-CNNs are employed to simultaneously extract spatial features from each frame, and temporal features across multiple frames, which are typically stacked along the 3<sup>rd</sup> (depth) dimension. The idea of exploiting 3D convolutions for videos was proposed in [32] for human action recognition, and was later extended to other applications [13]. **In self-supervised learning, however, the number of pretext tasks that exploit this technique is limited.** Kim *et al.* [33] proposed a task that extracts cubic puzzles of  $2 \times 2 \times 1$ , meaning that the 3<sup>rd</sup> dimension is not actually utilized in puzzle creation. Jing *et al.* [34] extended the rotation prediction task [19] to videos, by simply stacking video frames along the depth dimension, however, this dimension is not employed in the design of their task as only spatial rotations are considered. Han *et al.* proposed a dense encoding of spatio-temporal frame blocks to predict future scene representations recurrently, in conjunction with a curriculum training scheme to extend the predicted future. Similarly, the depth dimension is not employed in this task. On the other hand, in our more general versions of 3D Jigsaw puzzles and 3D Rotation prediction, respectively, we exploit the depth (3<sup>rd</sup>) dimension in the design of our tasks. For instance, we solve larger 3D puzzles up to  $3 \times 3 \times 3$ , and we also predict more rotations along all axes in the 3D space. **Furthermore, in our 3D Contrastive Predictive Coding task, we predict patch representations along all 3 dimensions, scanning input volumes in a manner that resembles a pyramid.** In general, we believe the different nature of the data, 3D volumetric scans vs. stacked video frames, influences the design of proxy tasks, i.e. the depth dimension has an actual semantic meaning in volumetric scans. Hence, we consider the whole 3D context when designing all of our methods, aiming to learn valuable anatomical information from unlabeled 3D volumetric scans.

In the medical context, self-supervision has found use-cases in diverse applications such as depth estimation in monocular endoscopy [35], robotic surgery [36], medical image registration [37], body part recognition [38], in disc degeneration using spinal MRIs [39], in cardiac image segmentation [40], body part regression for slice ordering [41], and medical instrument segmentation [42]. Spitzer *et al.* [43] **sample 2D patches from a 3D brain, and predict the distance between these patches as a supervision signal.** Tajbakhsh *et al.* [44] use orientation prediction from medical images as a proxy task. There are multiple other examples of self-supervised methods for medical imaging, such as [45–49]. While these attempts are a step forward for self-supervised learning in medical imaging, they have some limitations. First, as opposed to our work, many of these works make assumptions about input data, resulting in engineered solutions that hardly generalize to other target tasks. Second, none of the above works capture the complete spatial context available in 3-dimensional scans, i.e. they only operate on 2D/2.5D spatial context. In a more related work, Zhou *et al.* [50] extended image reconstruction techniques from 2D to 3D, and implemented multiple self-supervised tasks based on image-reconstruction. Zhuang *et al.* [51] and Zhu *et al.* [52] developed a proxy task that solves small 3D jigsaw puzzles. Their proposed puzzles were only limited to  $2 \times 2 \times 2$  of puzzle complexity. **Our version of 3D Jigsaw puzzles is able to efficiently solve larger puzzles, e.g.  $3 \times 3 \times 3$ , and outperforms their method’s results on the downstream task of Brain tumor segmentation.** In this paper, we continue this line of work, and develop five different algorithms for 3D data, whose nature and performance can accommodate more types of target medical applications.

### 3 Self-Supervised Methods

In this section, we discuss the formulations of our 3D self-supervised pretext tasks, all of which learn data representations from unlabeled samples (3D images), hence requiring no manual annotation effort in the self-supervised pretraining stage. Each task results in a **pretrained encoder model  $g_{enc}$**  that can be fine-tuned in various downstream tasks, subsequently.

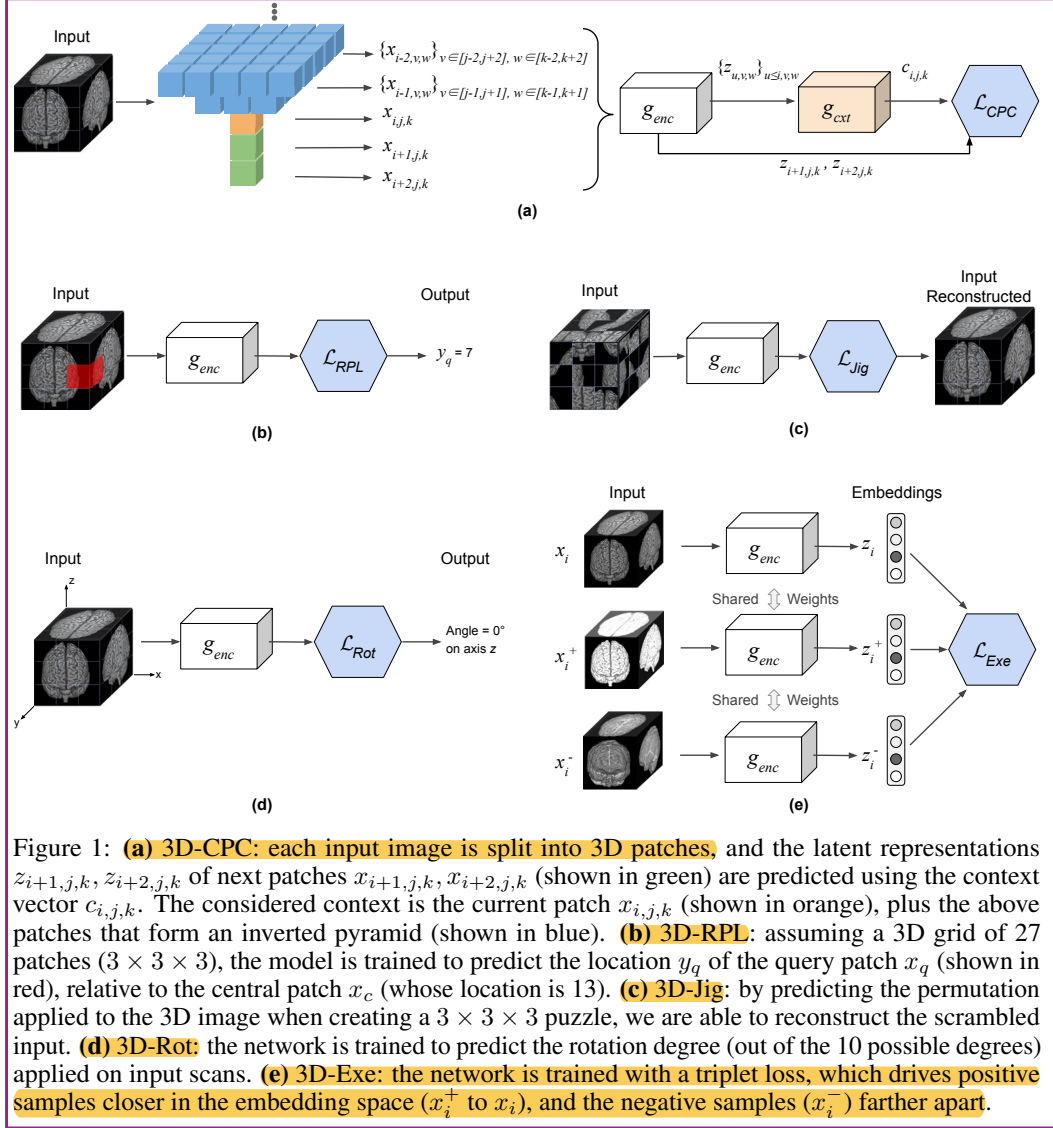


Figure 1: **(a) 3D-CPC**: each input image is split into 3D patches, and the latent representations  $z_{i+1,j,k}, z_{i+2,j,k}$  of next patches  $x_{i+1,j,k}, x_{i+2,j,k}$  (shown in green) are predicted using the context vector  $c_{i,j,k}$ . The considered context is the current patch  $x_{i,j,k}$  (shown in orange), plus the above patches that form an inverted pyramid (shown in blue). **(b) 3D-RPL**: assuming a 3D grid of 27 patches ( $3 \times 3 \times 3$ ), the model is trained to predict the location  $y_q$  of the query patch  $x_q$  (shown in red), relative to the central patch  $x_c$  (whose location is 13). **(c) 3D-Jig**: by predicting the permutation applied to the 3D image when creating a  $3 \times 3 \times 3$  puzzle, we are able to reconstruct the scrambled input. **(d) 3D-Rot**: the network is trained to predict the rotation degree (out of the 10 possible degrees) applied on input scans. **(e) 3D-Exe**: the network is trained with a triplet loss, which drives positive samples closer in the embedding space ( $x_i^+$  to  $x_i$ ), and the negative samples ( $x_i^-$ ) farther apart.

### 3.1 3D Contrastive Predictive Coding (3D-CPC)

Following the contrastive learning idea, first proposed in [26], this universal unsupervised technique predicts the latent space for future (next or adjacent) samples. Recently, CPC found success in multiple application fields, e.g. its 1D version in audio signals [22], and its 2D versions in images [22][23], and was able to bridge the gap between unsupervised and fully-supervised methods [24]. Our proposed CPC version generalizes this technique to 3D inputs, and defines a proxy task by cropping equally-sized and overlapping 3D patches from each input scan. Then, the encoder model  $g_{enc}$  maps each input patch  $x_{i,j,k}$  to its latent representation  $z_{i,j,k} = g_{enc}(x_{i,j,k})$ . Next, another model called the context network  $g_{ctx}$  is used to summarize the latent vectors of the patches in the context of  $x_{i,j,k}$ , and produce its context vector  $c_{i,j,k} = g_{ctx}(\{z_{u,v,w}\}_{u \leq i,v,w})$ , where  $\{z\}$  denotes a set of latent vectors. Finally, because  $c_{i,j,k}$  captures the high level content of the context that corresponds to  $x_{i,j,k}$ , it allows for predicting the latent representations of next (adjacent) patches  $z_{i+l,j,k}$ , where  $l \geq 0$ . This prediction task is cast as an  $N$ -way classification problem by utilizing the InfoNCE loss [22], which takes its name from its ability to maximize the mutual information between  $c_{i,j,k}$  and  $z_{i+l,j,k}$ . Here, the classes are the latent representations  $\{z\}$  of the patches, among which is one positive representation, and the rest  $N - 1$  are negative. Formally, the CPC loss can be written as

follows:

$$\begin{aligned}\mathcal{L}_{CPC} &= - \sum_{i,j,k,l} \log p(z_{i+l,j,k} \mid \hat{z}_{i+l,j,k}, \{z_n\}) \\ &= - \sum_{i,j,k,l} \log \frac{\exp(\hat{z}_{i+l,j,k} z_{i+l,j,k})}{\exp(\hat{z}_{i+l,j,k} z_{i+l,j,k}) + \exp(\sum_n \hat{z}_{i+l,j,k} z_n)}\end{aligned}\quad (1)$$

This loss corresponds to the categorical cross-entropy loss, which trains the model to recognize the correct representation  $z_{i+l,j,k}$  among the list of negative representations  $\{z_n\}$ . These negative samples (3D patches) are chosen randomly from other locations in the input image. In practice, similar to the original NCE [26], this task is solved as a binary pairwise classification task.

It is noteworthy that the proposed 3D-CPC task, illustrated in Fig. 1(a), allows employing any network architecture in the encoder  $g_{enc}$  and the context  $g_{ctx}$  networks. In our experiments, we follow [22] in using an autoregressive network using GRUs [53] for the context network  $g_{ctx}$ , however, masked convolutions can be a valid alternative [54]. In terms of what the 3D context of each patch  $x_{i,j,k}$  includes, we follow the idea of an inverted pyramid neighborhood, which is inspired from [55][56]. This context is chosen based on a tradeoff between computational cost and performance. Too large contexts (e.g. full surrounding of a patch) incur prohibitive computations and memory use. The inverted-pyramid context was an optimal tradeoff.

### 3.2 Relative 3D patch location (3D-RPL)

In this task, the spatial context in images is leveraged as a rich source of supervision, in order to learn semantic representations of the data. First proposed by Doersch *et al.* [15] for 2D images, this task inspired several works in self-supervision. In our 3D version, shown in Fig. 1(b), we leverage the full 3D spatial context in the design of our task. From each input 3D image, a 3D grid of  $N$  non-overlapping patches  $\{x_i\}_{i \in \{1, \dots, N\}}$  is sampled at random locations. Then, the patch  $x_c$  in the center of the grid is used as a reference, and a query patch  $x_q$  is selected from the surrounding  $N - 1$  patches. Next, the location of  $x_q$  relative to  $x_c$  is used as the positive label  $y_q$ . This casts the task as an  $N - 1$ -way classification problem, in which the locations of the remaining grid patches are used as the negative samples  $\{y_n\}$ . Formally, the cross-entropy loss in this task is written as:

$$\mathcal{L}_{RPL} = - \sum_{k=1}^K \log p(y_q \mid \hat{y}_q, \{y_n\}) \quad (2)$$

Where  $K$  is the number of queries extracted from all samples. In order to prevent the model from solving this task quickly by finding shortcut solutions, e.g. edge continuity, we follow [15] in leaving random gaps (jitter) between neighboring 3D patches. More details in Appendix.

### 3.3 3D Jigsaw puzzle Solving (3D-Jig)

Deriving a Jigsaw puzzle grid from an input image, be it in 2D or 3D, and solving it can be viewed as an extension to the above patch-based RPL task. In our 3D Jigsaw puzzle task, which is inspired by its 2D counterpart [16] and illustrated in Fig. 1(c), the puzzles are formed by sampling an  $n \times n \times n$  grid of 3D patches. Then, these patches are shuffled according to an arbitrary permutation, selected from a set of predefined permutations. This set of permutations with size  $P$  is chosen out of the  $n^3!$  possible permutations, by following the Hamming distance based algorithm in [16] (details in Appendix), and each permutation is assigned an index  $y_p \in \{1, \dots, P\}$ . Therefore, the problem is cast as a  $P$ -way classification task, i.e., the model is trained to simply recognize the applied permutation index  $p$ , allowing us to solve the 3D puzzles in an efficient manner. Formally, we minimize the cross-entropy loss of  $\mathcal{L}_{Jig}(y_p^k, \hat{y}_p^k)$ , where  $k \in \{1, \dots, K\}$  is an arbitrary 3D puzzle from the list of extracted  $K$  puzzles. Similar to 3D-RPL, we use the trick of adding random jitter in 3D-Jig.

### 3.4 3D Rotation prediction (3D-Rot)

Originally proposed by Gidaris *et al.* [19], the rotation prediction task encourages the model to learn visual representations by simply predicting the angle by which the input image is rotated. The intuition behind this task is that for a model to successfully predict the angle of rotation, it needs



to capture sufficient semantic information about the object in the input image. In our 3D Rotation prediction task, 3D input images are rotated randomly by a random degree  $r \in \{1, \dots, R\}$  out of the  $R$  considered degrees. In this task, for simplicity, we consider the multiples of 90 degrees ( $0^\circ, 90^\circ, 180^\circ, 270^\circ$ , along each axis of the 3D coordinate system  $(x, y, z)$ ). There are 4 possible rotations per axis, amounting to 12 possible rotations. However, rotating input scans by  $0^\circ$  along the 3 axes will produce 3 identical versions of the original scan, hence, we consider 10 rotation degrees instead. Therefore, in this setting, this proxy task can be solved as a 10-way classification problem. Then, the model is tasked to predict the rotation degree (class), as shown in Fig. 1(d). Formally, we minimize the cross-entropy loss  $\mathcal{L}_{Rot}(r^k, \hat{r}^k)$ , where  $k \in \{1, \dots, K\}$  is an arbitrary rotated 3D image from the list of  $K$  rotated images. It is noteworthy that we create multiple rotated versions for each 3D image.

### 3.5 3D Exemplar networks (3D-Exe)

The task of Exemplar networks, proposed by Dosovitskiy *et al.* [57], is one of the earliest methods in the self-supervised family. To derive supervision labels, it relies on image augmentation techniques, i.e. transformations. Assuming a training set  $X = \{x_1, \dots, x_N\}$ , and a set of  $K$  image transformations  $\mathcal{T} = \{T_1, \dots, T_K\}$ , a new surrogate class  $S_{x_i}$  is created by transforming each training sample  $x_i \in X$ , where  $S_{x_i} = \mathcal{T}x_i = \{Tx_i \mid T \in \mathcal{T}\}$ . Therefore, the task is cast as a regular classification task with a cross-entropy loss. However, this classification task becomes prohibitively expensive as the dataset size grows larger, as the number of classes grows accordingly. Thus, in our proposed 3D version of Exemplar networks, shown in Fig. 1(e), we employ a different mechanism that relies on the triplet loss instead [58]. Formally, assuming  $x_i$  is a random training sample and  $z_i$  is its corresponding embedding vector,  $x_i^+$  is a transformed version of  $x_i$  (seen as a positive example) with an embedding  $z_i^+$ , and  $x_i^-$  is a different sample from the dataset (seen as negative) with an embedding  $z_i^-$ . The triplet loss is written as follows:

$$\mathcal{L}_{Exe} = \frac{1}{N_T} \sum_{i=1}^{N_T} \max\{0, D(z_i, z_i^+) - D(z_i, z_i^-) + \alpha\} \quad (3)$$

where  $D(\cdot)$  is a pairwise distance function, for which we use the  $L_2$  distance, following [59].  $\alpha$  is a margin (gap) that is enforced between positive and negative pairs, which we set to 1. The triplet loss enforces  $D(z_i, z_i^-) > D(z_i, z_i^+)$ , i.e. the transformed versions of the same sample (positive samples) to come closer to each other in the learned embedding space, and farther away from other (negative) samples. Replacing the triplet loss with a contrastive loss [26] is possible in this method, and has been found to improve learned representations from natural images [24]. In addition, the learned representations by Exemplar can be affected by the negatives sampling strategy. The simple option is to sample from within the same batch, however, it is also possible to sample from the whole dataset. The latter choice is computationally more expensive, but is expected to improve the learned representations, as it makes the task harder. It is noteworthy that we apply the following 3D transformations: random flipping along an arbitrary axis, random rotation along an arbitrary axis, random brightness and contrast, and random zooming.

## 4 Experimental Results

In this section, we present the evaluation results of our methods, which we assess the quality of their learned representations by fine-tuning them on three downstream tasks. In each task, we analyze the obtained gains in data-efficiency, performance, and speed of convergence. In addition, each task aims to demonstrate a certain use-case for our methods. We follow the commonly used evaluation protocols for self-supervised methods in each of these tasks. The chosen tasks are:

- **Brain Tumor Segmentation from 3D MRI** (Subsection 4.1): in which we study the possibility for transfer learning from a different unlabeled 3D corpus, following [60].
- **Pancreas Tumor Segmentation from 3D CT** (Subsection 4.2): to demonstrate how to use the same unlabeled dataset, following the data-efficient evaluation protocol in [23].
- **Diabetic Retinopathy Detection from 2D Fundus Images** (Subsection 4.3): to showcase our implementations for the 2D versions of our methods, following [23]. Here, we also evaluate pretraining on a different large corpus, then fine-tuning on the downstream dataset.

We provide additional details about architectures, training procedures, the effect of augmentation in Exemplar, and how we initialize decoders for segmentation tasks in the Appendix.

#### 4.1 Brain Tumor Segmentation Results

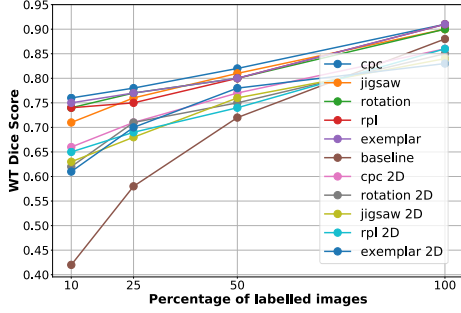
In this task, we evaluate our methods by fine-tuning the learned representations on the Multimodal Brain Tumor Segmentation (BraTS) 2018 [61, 62] benchmark. Before that, we pretrain our models on brain MRI data from the UK Biobank [63] (UKB) corpus, which contains roughly 22K 3D scans. Due to this large number of unlabeled scans, UKB is suitable for unsupervised pretraining. The BraTS dataset contains annotated MRI scans for 285 training and 66 validation cases. We fine-tune on BraTS’ training set, and evaluate on its validation set. Following the official BraTS challenge, we report Dice scores for the Whole Tumor (WT), Tumor Core (TC), and Enhanced Tumor (ET) tasks. The Dice score (F1-Score) is twice the area of overlap between two segmentation masks divided by the total number of pixels in both. In order to assess the quality of the learned representations by our 3D proxy tasks, we compare to the following baselines:

- Training from scratch: the first sensible baseline for any self-supervised method, in general, is the same model trained on the downstream task when initialized from random weights. Comparing to this baseline provides insights about the benefits of self-supervised pretraining.
- Training on 2D slices: this baseline aims to quantitatively show how our proposal to operate on the 3D context benefits the learned representations, compared to 2D methods.
- Supervised pretraining: this baseline uses automatic segmentation labels from FSL-FAST [64], which include masks for three brain tissues.
- Baselines from the BraTS challenge: we compare to the methods [65–68], which all use a single model with an architecture similar to ours, i.e. 3D U-Net [69].

**Discussion.** We first assess the gains in data-efficiency in this task. To quantify these gains, we measure the segmentation performance at different sample sizes. We randomly select subsets of patients at 10%, 25%, 50%, and 100% of the full dataset size, and we fine-tune our models on these subsets. Here, we compare to the baselines listed above. As shown in Fig. 2, our 3D methods outperform the baseline model trained from scratch by a large margin when using few training samples, and behaves similarly as the number of labeled samples increases. The low-data regime case at 5% suggests the potential for generic unsupervised features, and highlights the huge gains in data-efficiency. Also, the proposed 3D versions considerably outperform their 2D counterparts, which are trained on slices extracted from the 3D images. We also measure how our methods affect the final brain tumor segmentation performance, in Table 1. All our methods outperform the baseline trained from scratch as well as their 2D counterparts, confirming the benefits of pretraining with our 3D tasks on downstream performance. We also achieve comparable results to baselines from the BraTS challenge, and we outperform these baselines in some cases, e.g. our 3D-RPL method outperforms all baselines in terms of ET and TC dice scores. Also, our model pretrained with 3D-Exemplar, with fewer downstream training epochs, matches the result of Isensee *et al.* [65] in terms of WT dice score, which is one of the top results on the BraTS 2018 challenge. In comparison to the supervised baseline using automatic FAST labels, we find that our results are comparable, outperforming this baseline in some cases. Our results in this downstream task also demonstrate the generalization ability of our 3D tasks across different domains. This result is significant, because medical datasets are supervision-starved, e.g. images may be collected as part of clinical routine, but much fewer (high-quality) labels are produced, due to annotation costs.

#### 4.2 Pancreas Tumor Segmentation Results

In this downstream task, we evaluate our models on 3D CT scans of Pancreas tumor from the medical decathlon benchmarks [70]. The Pancreas dataset contains annotated CT scans for 420 cases. Each scan in this dataset contains 3 different classes: pancreas (class 1), tumor (class 2), and background (class 0). To measure the performance on this benchmark, two dice scores are computed for classes 1 and 2. In this task, we pretrain using our proposed 3D tasks on pancreas scans *without* their annotation masks. Then, we fine-tune the obtained models on subsets of annotated data to assess the gains in both data-efficiency and performance. Finally, we also compare to the baseline model trained from scratch and to 2D models, similar to the previous downstream task. Fig. 3 demonstrates the gains



**Figure 2:** Data-efficient segmentation results in BraTS. With less labeled data, the supervised baseline (brown) fails to generalize, as opposed to our methods. Also, the proposed 3D methods outperform all 2D counterparts.

**Table 1:** BraTS segmentation results

Model	ET	WT	TC
3D-From scratch	76.38	87.82	83.11
3D Supervised	78.88	90.11	84.92
2D-CPC	76.60	86.27	82.41
2D-RPL	77.53	87.91	82.56
2D-Jigsaw	76.12	86.28	83.26
2D-Rotation	76.60	88.78	82.41
2D-Exemplar	75.22	84.82	81.87
Popli <i>et al.</i> [66]	74.39	89.41	82.48
Baid <i>et al.</i> [67]	74.80	87.80	82.66
Chandra <i>et al.</i> [68]	74.06	87.19	79.89
Isensee <i>et al.</i> [65]	80.36	<b>90.80</b>	84.32
3D-CPC	80.83	89.88	85.11
3D-RPL	<b>81.28</b>	90.71	<b>86.12</b>
3D-Jigsaw	79.66	89.20	82.52
3D-Rotation	80.21	89.63	84.75
3D-Exemplar	79.46	<b>90.80</b>	83.87

when fine-tuning our models on 5%, 10%, 50%, and 100% of the full data size. The results obtained by our 3D methods also outperform the baselines in this task with a margin when using only few training samples, e.g. 5% and 10% cases. Another significant benefit offered by pretraining with our methods is the speed of convergence on downstream tasks. As demonstrated in Fig 5 when training on the full pancreas dataset, within the **first 20 epochs only**, our models achieve much higher performances compared to the "from scratch" baseline. We should note that we evaluate this task on a held-out labeled subset of the Pancreas dataset that was not used for pretraining nor fine-tuning. We provide the full list of experimental results for this task in Appendix.

### 4.3 Diabetic Retinopathy Results

As part of our work, we also provide implementations for the 2D versions of the developed self-supervised methods. **We showcase these implementations on the Diabetic Retinopathy 2019 Kaggle challenge [71].** This dataset **contains roughly 5590 Fundus 2D images**, each of which was rated by a clinician on a **severity scale of 0 to 4**. These levels define a classification task. In order to evaluate our tasks on this benchmark, we pretrain all the 2D versions of our methods using 2D Fundus images from UK Biobank [63]. **The retinopathy data in UK Biobank contains 170K images.** We then fine-tune the obtained models on Kaggle data, meaning performing transfer learning. We also compare the obtained results with this transfer learning protocol to those obtained with the data-efficient evaluation protocol in [23], i.e. pretraining on the same Kaggle dataset and fine-tuning on subsets of it. To assess the gains in data-efficiency, we fine-tune the obtained models on subsets of labelled Kaggle data, shown in Fig. 4. It is noteworthy that pretraining on UKB produces results that outperform those obtained when pretraining on the same Kaggle dataset. This confirms the benefits of transfer learning from a large corpus to a smaller one using our methods. Gains in speed of convergence are also shown in Fig. 6. In this 2D task, we achieve results consistent with the other downstream tasks, presented before. We should point out that we evaluate with 5-fold cross validation on this 2D dataset. The metric used in task, as in the Kaggle challenge, is the Quadratic Weighted Kappa, which measures the agreement between two ratings. Its values vary from random (0) to complete (1) agreement, and if there is less agreement than chance it may become negative.

## 5 Conclusion

In this work, we asked whether designing 3D self-supervised tasks could benefit the learned representations from unlabeled 3D images, and found that it indeed greatly improves their downstream performance, especially when fine-tuned on only small amounts of labeled 3D data. **We demonstrate the obtained gains by our proposed 3D algorithms in data-efficiency, performance, and speed of convergence on three different downstream tasks. Our 3D tasks outperform their 2D counterparts, hence supporting our proposal of utilizing the 3D spatial context in the design of self-supervised tasks, when operating on 3D domains.** What is more, our results, particularly in the low-data regime,



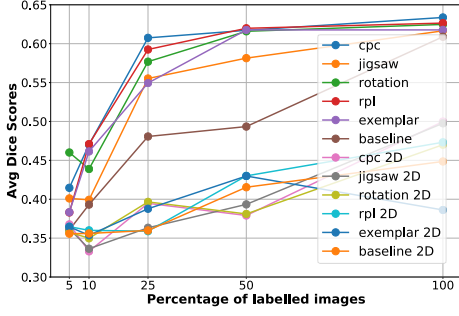


Figure 3: Data-efficient segmentation results in Pancreas. With less labeled data, the supervised baseline (brown) fails to generalize, as opposed to our methods. Also, the proposed 3D methods outperform all 2D counterparts

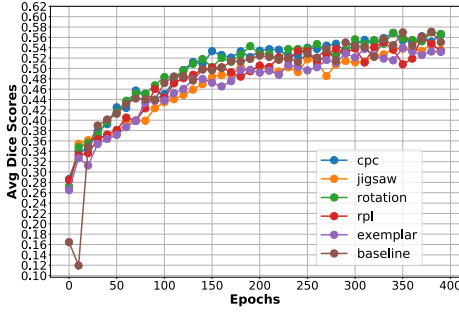


Figure 5: Speed of convergence in Pancreas segmentation. Our models converge faster than the baseline (brown)

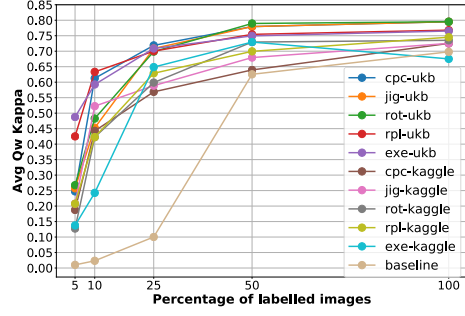


Figure 4: Data-efficient classification in Diabetic Retinopathy. With less labels, the supervised baseline (brown) fails to generalize, as opposed to pretrained models. This result is consistent with the other downstream tasks

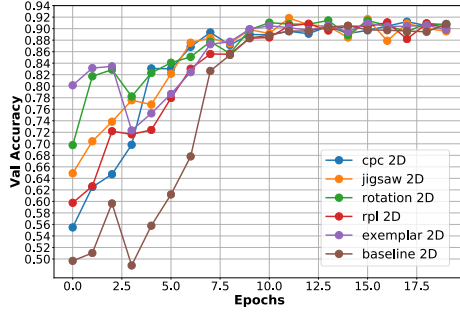


Figure 6: Speed of convergence in Retinopathy classification. Our models also converge faster in this task

demonstrate the possibility to reduce the manual annotation effort required in the medical imaging domain, where data and annotation scarcity is an obstacle. Furthermore, we observe performance gains when pretraining our methods on a large unlabeled corpus, and fine-tuning them on a different smaller downstream-specific dataset. **This result suggests alternatives for transfer learning from Imagenet features, which can be substantially different from the medical domain.** Finally, we open source our implementations for all 3D methods (and also their 2D versions), and we publish them to help other researchers apply our methods on other medical imaging tasks. This work is only a first step toward creating a set of methods that facilitate self-supervised learning research for 3D data, e.g. medical scans. We believe there is room for improvement along this line, such as designing new 3D proxy tasks, evaluating different architectural options, and including other data modalities (e.g. text) in conjunction with images/scans.

## Broader Impact

Due to technological advancements in 3D data sensing, and to the growing number of its applications, the attention to machine learning algorithms that perform analysis tasks on such data has grown rapidly in the past few years. As mentioned before, 3D imaging has multitude of applications [2], such as in Robotics, in CAD imaging, in Geology, and in Medical Imaging. In this work, we developed multiple 3D Deep Learning algorithms, and evaluated them on multiple 3D medical imaging benchmarks. Our focus on medical imaging is motivated by the pressing demand for automatic (and instant) analysis systems, that may aid the medical community.

Medical imaging plays an important role in patient healthcare, as it aids in disease prevention, early detection, diagnosis, and treatment. With the continuous digitization of medical images, the hope that physicians and radiologists are able to instantly analyze them with Machine Learning algorithms is slowly shaping as a reality. Achieving this has become more critical recently, as the number of patients which contracted with a novel Coronavirus, called COVID-19, reached a high record. Radiography images provide a rich and a quick diagnosis tool, because other types of tests, e.g. RT-PCR which is an RNA/DNA based test, have low sensitivity and may require hours/days of processing [72]. Therefore, as imaging allows such instant insights into human body organs, it receives growing attention from both machine learning and medical communities.

Yet efforts to leverage advancements in machine learning, particularly the supervised algorithms, are often hampered by the sheer expense of expert annotation required [4]. Generating expert annotations of patient data at scale is non-trivial, expensive, and time-consuming, especially for 3D medical scans. Even current semi-automatic software tools fail to sufficiently address this challenge. Consequently, it is necessary to rely on annotation-efficient machine learning algorithms, such as self-supervised (unsupervised) approaches for representation learning from unlabelled data. Our work aims to provide the necessary tools for 3D image analysis, in general, and to aid physicians and radiologists in their diagnostic tasks from 3D scans, in particular. And as the main consequence of this work, the developed methods can help reduce the effort and cost of annotation required by these practitioners. In the larger goal of leveraging Machine Learning for good, our work is only a small step toward achieving this goal for patient healthcare.

## Acknowledgments and Disclosure of Funding

This research has been supported by funding from the German Federal Ministry of Education and Research (BMBF) in the project KI-LAB-ITSE (project number 01IS19066). This research has been conducted using the UK Biobank Resource.

## References

- [1] David Griffiths and Jan Boehm. A review on deep learning techniques for 3d sensed data classification. *CoRR*, abs/1907.04444, 2019. URL <http://arxiv.org/abs/1907.04444>
- [2] Anastasia Ioannidou, Elisavet Chatzilari, Spiros Nikolopoulos, and Ioannis Kompatsiaris. Deep learning advances in computer vision with 3d data: A survey. *ACM Computing Surveys*, 50, 06 2017. doi: 10.1145/3042064.
- [3] Hao Su, Leonidas Guibas, Michael Bronstein, Evangelos Kalogerakis, Jimei Yang, Charles Qi, and Qixing Huang. *3D Deep Learning*, 2017 (accessed June 2, 2020). URL <http://3ddl.stanford.edu/>
- [4] Katharina Grünberg, Oscar Jimenez-del Toro, Andras Jakab, Georg Langs, Tomàs Salas Fernandez, Marianne Winterstein, Marc-André Weber, and Markus Krenn. *Annotating Medical Image Data*, pages 45–67. Springer International Publishing, Cham, 2017.
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR09*, Miami, FL, USA, 2009. IEEE.
- [6] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

- [7] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017. URL <http://arxiv.org/abs/1711.05225>
- [8] Jaakko Sahlsten, Joel Jaskari, Jyri Kivinen, Lauri Turunen, Esa Jaanio, Kustaa Hietala, and Kimmo Kaski. Deep learning fundus image analysis for diabetic retinopathy and macular edema grading. *Scientific Reports*, 9, 12 2019. doi: 10.1038/s41598-019-47181-w.
- [9] Sheikh Muhammad Saiful Islam, Md Mahedi Hasan, and Sohaib Abdullah. Deep learning based early detection and grading of diabetic retinopathy using retinal fundus images. *CoRR*, abs/1812.10595, 2018. URL <http://arxiv.org/abs/1812.10595>
- [10] Antonio Torralba and Alexey A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011.
- [11] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems 32*, pages 3347–3357. Curran Associates, Inc., 2019. URL <http://papers.nips.cc/paper/8596-transfusion-understanding-transfer-learning-for-medical-imaging.pdf>
- [12] Ronald Eisenberg and Alexander Margulis. *A Patient’s Guide to Medical Imaging*. New York: Oxford University Press, NY, USA, 2011.
- [13] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *CoRR*, abs/1902.06162, 2019. URL <http://arxiv.org/abs/1902.06162>
- [14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, May 2-4, 2013, Workshop Track Proceedings*, Scottsdale, Arizona, USA, 2013. OpenReview. URL <http://arxiv.org/abs/1301.3781>
- [15] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV ’15*, page 1422–1430, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.167. URL <https://doi.org/10.1109/ICCV.2015.167>
- [16] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 69–84, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46466-4.
- [17] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 649–666, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46487-9.
- [18] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *The European Conference on Computer Vision (ECCV)*, Munich, Germany, September 2018. Springer.
- [19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *CoRR*, abs/1803.07728, 2018. URL <http://arxiv.org/abs/1803.07728>
- [20] Jiawei Wang, Shuai Zhu, Jiao Xu, and Da Cao. The retrieval of the beautiful: Self-supervised salient object detection for beauty product retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia, MM ’19*, page 2548–2552, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3356059. URL <https://doi.org/10.1145/3343031.3356059>
- [21] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [22] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL <http://arxiv.org/abs/1807.03748>
- [23] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. *CoRR*, abs/1905.09272, 2019. URL <http://arxiv.org/abs/1905.09272>

- [24] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.
- [25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [26] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterton, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL <http://proceedings.mlr.press/v9/gutmann10a.html>
- [27] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2015.
- [28] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating the future by watching unlabeled video. *CoRR*, abs/1504.08023, 2015. URL <http://arxiv.org/abs/1504.08023>.
- [29] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2443–2451, 2015.
- [30] Senthil Purushwalkam and Abhinav Gupta. Pose from action: Unsupervised learning of pose features based on motion. *CoRR*, abs/1609.05420, 2016. URL <http://arxiv.org/abs/1609.05420>
- [31] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [32] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.
- [33] Kim Dahun, Donghyeon Cho, and Soo-Ok Kweon. Self-supervised video representation learning with space-time cubic puzzles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8545–8552, 07 2019. doi: 10.1609/aaai.v33i01.33018545.
- [34] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *CoRR*, abs/1811.11387, 2018. URL <http://arxiv.org/abs/1811.11387>
- [35] Xingtong Liu, Ayushi Sinha, Mathias Unberath, Masaru Ishii, Gregory D. Hager, Russell H. Taylor, and Austin Reiter. Self-supervised learning for dense depth estimation in monocular endoscopy. *CoRR*, abs/1806.09521, 2018. URL <http://arxiv.org/abs/1806.09521>
- [36] Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang Yang. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. In *The Hamlyn Symposium on Medical Robotics*, pages 27–28, 06 2017. doi: 10.31256/HSMR2017.14.
- [37] Hongming Li and Yong Fan. Non-rigid image registration using self-supervised fully convolutional networks without training data. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1075–1078, Washington, DC, USA, April 2018. IEEE.
- [38] Pengyue Zhang, Fusheng Wang, and Yefeng Zheng. Self supervised deep representation learning for fine-grained body part recognition. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 578–582, Melbourne, Australia, April 2017. IEEE.
- [39] Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Self-supervised learning for spinal mris. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 294–302, Cham, 09 2017. Springer. ISBN 978-3-319-67557-2. doi: 10.1007/978-3-319-67558-9\_34.
- [40] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E. Petersen, Yike Guo, Paul M. Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 541–549, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32245-8.

- [41] Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam P. Harrison, Mohammadhadi Bagheri, and Ronald M. Summers. *Deep Lesion Graph in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-Scale Lesion Database*, pages 413–435. Springer International Publishing, Cham, 2019. ISBN 978-3-030-13969-8. doi: 10.1007/978-3-030-13969-8\_20. URL [https://doi.org/10.1007/978-3-030-13969-8\\_20](https://doi.org/10.1007/978-3-030-13969-8_20)
- [42] Tobias Roß, David Zimmerer, Anant Vemuri, Fabian Isensee, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Müller, Hannes Kenngott, Stefanie Speidel, Klaus Maier-Hein, and Lena Maier-Hein. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International Journal of Computer Assisted Radiology and Surgery*, 13, 11 2017. doi: 10.1007/s11548-018-1772-0.
- [43] Hannah Spitzer, Kai Kiwitz, Katrin Amunts, Stefan Harmeling, and Timo Dickscheid. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 663–671, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00931-1.
- [44] N. Tajbakhsh, Y. Hu, J. Cao, X. Yan, Y. Xiao, Y. Lu, J. Liang, D. Terzopoulos, and X. Ding. 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)*, pages 1251–1255, 2019.
- [45] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 2019. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2019.101539>. URL <http://www.sciencedirect.com/science/article/pii/S1361841518304699>
- [46] Jianbo Jiao, Richard Droste, Lior Drukker, Aris T. Papageorgiou, and J. Alison Noble. Self-supervised representation learning for ultrasound video. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1847–1850, 2020.
- [47] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis, 2019.
- [48] Maximilian Blendowski, Hannes Nickisch, and Matthias P. Heinrich. How to learn from unlabeled volume data: Self-supervised 3d context feature learning. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 649–657, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32226-7.
- [49] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations, 2020.
- [50] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 384–393, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32251-9.
- [51] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik’s cube. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 420–428, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32251-9.
- [52] Jiuwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S. Kevin Zhou, and Yefeng Zheng. Rubik’s cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical Image Analysis*, 64:101746, 2020. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2020.101746>. URL <http://www.sciencedirect.com/science/article/pii/S1361841520301109>
- [53] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL <https://www.aclweb.org/anthology/D14-1179>



- [54] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4790–4798. Curran Associates, Inc., 2016. URL <http://papers.nips.cc/paper/6527-conditional-image-generation-with-pixelcnn-decoders.pdf>
- [55] Marijn F. Stollenga, Wonmin Byeon, Marcus Liwicki, and Jürgen Schmidhuber. Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation. *CoRR*, abs/1506.07452, 2015. URL <http://arxiv.org/abs/1506.07452>
- [56] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML’16, page 1747–1756. JMLR.org, 2016.
- [57] Alexey Dosovitskiy, Jost T. Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014. URL <http://lmb.informatik.uni-freiburg.de/Publications/2014/DB14b>
- [58] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2015.
- [59] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.
- [60] Priya Goyal, Dhruv Mahajan, Harikrishna Mulam, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 6390–6399, October 2019. doi: 10.1109/ICCV.2019.00649.
- [61] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, and et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.
- [62] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4:170117 EP –, 09 2017.
- [63] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):1–10, 03 2015. doi: 10.1371/journal.pmed.1001779. URL <https://doi.org/10.1371/journal.pmed.1001779>
- [64] Mark W. Woolrich, Saad Jbabdi, Brian Patenaude, Michael Chappell, Salima Makni, Timothy Behrens, Christian Beckmann, Mark Jenkinson, and Stephen M. Smith. Bayesian analysis of neuroimaging data in fsl. *NeuroImage*, 45(1, Supplement 1):S173 – S186, 2009. ISSN 1053-8119. doi: <https://doi.org/10.1016/j.neuroimage.2008.10.055>. URL <http://www.sciencedirect.com/science/article/pii/S1053811908012044> Mathematics in Brain Imaging.
- [65] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. No new-net. In *International MICCAI Brainlesion Workshop*, pages 234–244, Granada, Spain, 2018. Springer.
- [66] Anmol Popli, Manu Agarwal, and G.N. Pillai. Automatic brain tumor segmentation using u-net based 3d fully convolutional network. In *Pre-Conference Proceedings of the 7th MICCAI BraTS Challenge*, pages 374–382. Springer, 2018.
- [67] Ujjwal Baid, Abhishek Mahajan, Sanjay Talbar, Swapnil Rane, Siddhesh Thakur, Aliasgar Moiyadi, Meenakshi Thakur, and Sudeep Gupta. Gbm segmentation with 3d u-net and survival prediction with radiomics. In *International MICCAI Brainlesion Workshop*, pages 28–35. Springer, 2018.
- [68] Siddhartha Chandra, Maria Vakalopoulou, Lucas Fidon, Enzo Battistella, Theo Estienne, Roger Sun, Charlotte Robert, Eric Deutch, and Nikos Paragios. Context aware 3-d residual networks for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 74–82. Springer, 2018.

- [69] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.
- [70] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert J. S. Litjens, Bjoern H. Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick Ferdinand Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan Heckers, William R. Jarnagin, Maureen McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *CoRR*, abs/1902.09063, 2019. URL <http://arxiv.org/abs/1902.09063>
- [71] Aptos 2019. <https://www.kaggle.com/c/aptos2019-blindness-detection/> 2019. Accessed: 2020-11-02.
- [72] Sayan Manna, Jill Wruble, Samuel Z Maron, Danielle Toussie, Nicholas Voutsinas, Mark Finkelstein, Mario A Cedillo, Jamie Diamond, Corey Eber, Adam Jacobi, Michael Chung, and Adam Bernheim. Covid-19: A multimodality review of radiologic techniques, clinical utility, and imaging features. *Radiology: Cardiothoracic Imaging*, 2(3):e200210, 2020. doi: 10.1148/ryct.2020200210. URL <https://doi.org/10.1148/ryct.2020200210>
- [73] tensorflow.org. *Tensorflow v2.1*, 2020 (accessed June 3, 2020). URL [https://www.tensorflow.org/versions/r2.1/api\\_docs/python/tf](https://www.tensorflow.org/versions/r2.1/api_docs/python/tf)
- [74] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.
- [75] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL <http://arxiv.org/abs/1412.6980> cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.
- [76] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [77] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA ’05, pages 399–402, New York, NY, USA, 2005. ACM. ISBN 1-59593-044-2. doi: 10.1145/1101149.1101236. URL <http://doi.acm.org/10.1145/1101149.1101236>

# Appendices

## A Implementation and training details for all tasks

It is noteworthy that our attached implementations are flexible enough to allow for evaluating several types of network architectures for encoders, decoders, and classifiers. We also provide implementations for multiple losses, augmentation techniques, and evaluation metrics. More information can be found in the README.md file in our attached code-base. We rely on tensorflow v2.1 [73] with Keras API in our implementations. Below, we provide the training details we used in implementing our 3D self-supervised tasks (and their 2D counterparts), and when fine-tuning them in subsequent downstream tasks.

**Architecture details.** For all 3D encoders  $g_{enc}$ , which are pretrained with our 3D self-supervised tasks and later fine-tuned on 3D segmentation tasks, we use a 3D U-Net [69]-based encoder (the downward path), which consists of five levels of residual convolutional blocks. The numbers of filters in these blocks are 32, 64, 128, 256, 512, respectively. The U-Net decoder (the upward path) is added in the downstream fine-tuning stage, and it includes five levels of deconvolutional blocks with skip connections from the U-Net encoder blocks. For the 2D encoders, we use a standard Densenet-121 [74] architecture, which is fine-tuned later on 2D classification tasks. When training our 3D self-supervised tasks, we follow [24] in adding nonlinear transformations (a hidden layer with ReLU activation) before the final classification layers. These classification layers are removed when fine-tuning the resulting encoders  $g_{enc}$  in downstream tasks.

**Optimization details.** In all self-supervised and downstream tasks, we use Adam [75] optimizer to train the models. The initial learning rate we use is 0.001 in 3D self-supervised tasks, 0.00001 in 3D segmentation tasks, 0.0005 in 2D self-supervised tasks, and 0.00005 in 2D classification tasks. When we fine-tune our pretrained encoders in subsequent downstream tasks, we follow a warm-up procedure inspired from [76] by keeping the encoder weights frozen for a number of initial warm-up epochs while the network decoders or classifiers are trained. These warm-up epochs are 5 in 2D classification tasks, and 25 epochs in 3D segmentation tasks. The alternative options we evaluated were: 1) fine-tuning the encoder directly with a randomly initialized decoder, 2) keeping the encoder frozen throughout the training procedure. And the 3<sup>rd</sup> option we followed in the end was the hybrid approach of warm-up epochs described above, as it provided a performance boost over the other alternatives. For segmentation tasks, in particular, where a decoder is used in the architecture, these warm-up epochs prove indispensable. Otherwise, training the whole model with a randomly initialized decoder, while the encoder is not frozen, may harm the encoder representations.

**Input preprocessing.** For all input scans, we perform the following preprocessing steps:

- In self-supervised pretraining using 3D scans, we find the boundaries of the brain or the pancreas along each axis, and then we crop the remaining empty parts from the scan. This step reduces the amount of empty background voxels, as they might confuse patch-based self-supervised methods with no additional semantic information. This step is not performed when fine-tuning on 3D downstream tasks.
- Then, we resize each 3D image from BraTS or Pancreas to a unified resolution of  $128 \times 128 \times 128$ , and to the resolution  $224 \times 224$  for 2D images from Diabetic Retinopathy.
- Then, each image’s intensity values are normalized by scaling them to the range  $[0, 1]$ .

**Processing multimodal inputs.** In the first downstream task of brain tumor segmentation with 3D multimodal MRI, we pretrain using the UK Biobank [63] corpus, as mentioned earlier. Brain scans obtained from UKB contain 2 MRI modalities (T1 and T2-Flair), which are co-registered. This allows us to stack these 2 modalities as color channels in each input sample, similar to RGB channels. This form of early fusion [77] of MRI modalities is common when they are registered, and is a practical solution for combining all information that exist in these modalities. However, as mentioned earlier, we use the BraTS [61, 62] dataset for fine-tuning, and each scan consists of 4 different MRI modalities, as opposed to only 2 in UKB that is used for pretraining. This difference only affects the input layer of the pretrained encoder, as fine-tuning on an incompatible number of input channels causes this process of fine-tuning to fail. We resolve this issue by duplicating (copying) the weights of *only* the pretrained input layer. This minor modification only adds a few additional parameters to the input layer, but allows us to leverage its weights. The other alternative for this solution would have been to discard the weights of this input layer, and initialize the rest of the model layers from pretrained models normally. But we believe our solution for this issue takes advantage of any useful information encoded in these weights. This multimodal inputs problem does not occur in the other downstream tasks, as the inputs include only one modality/channel.

### Task specific training details.

- **3D-CPC and 3D-Exe**: we use latent representation code size of 1024 in these tasks.
- **3D-Jig and 3D-RPL**: We split the input 3D images into  $3 \times 3 \times 3$  patches in this task. We apply a random jitter of 3 pixels per side (axis).
- **Patch-based tasks (3D-CPC, 3D-RPL, 3D-Jig)**: each extracted patch is represented using an embedding vector of size 64.
- **3D-Exe**: the  $\alpha$  value used for the triplet loss is 1.0.
- **3D-Jig**: the complexity of the Jigsaw puzzle solving task relies on the number of permutations used in generating the puzzles, i.e. the more permutations used, the harder the task is to solve. We follow the Hamming distance-based algorithm from [16] in sampling the permutations for this task. However, in our 3D puzzles task, we sample permutations that are more complex with 27 different entries. This algorithm works as follows: we sample a subset of 1000 permutations which are selected based on their Hamming distance, i.e., the number of different tile locations between 2 permutations. When generating permutations, we ensure that the average Hamming distance across permutations is kept as high as possible. This results in a set of permutations (classes) that are as far as possible from each other.

**Augmentation in Exemplar.** As mentioned earlier, we apply the following 3D transformations in Exemplar: random flipping along an arbitrary axis, random rotation along an arbitrary axis, random brightness and contrast, and random zooming. These augmentations are utilized to produce the positive samples. We vary the percentages of applying these augmentations using these factors:  $\alpha = 0.5$  for random rotations,  $\beta = 0.5$  for color distortions (brightness and contrast), and  $\gamma = 0.2$  for random zooming. When trying to omit a certain augmentation from the list above, we observe a drop in downstream performance. This is consistent with the findings of [24]. However, performing such transformations for high percentages is time-consuming, hence the reduced rates to 50%. Conducting a more thorough analysis of what *types* of augmentations are desirable is a future work.

## B Detailed experimental results

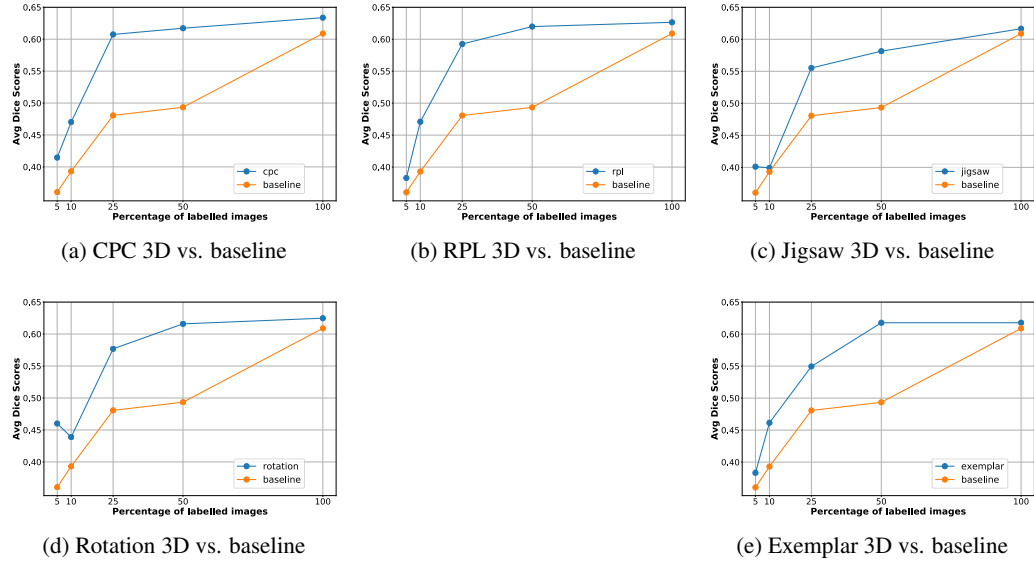


Figure 7: Pancreas segmentation: Detailed data-efficiency results per method (blue) vs. the supervised baseline (orange). Our methods consistently outperform the baseline in low-data cases

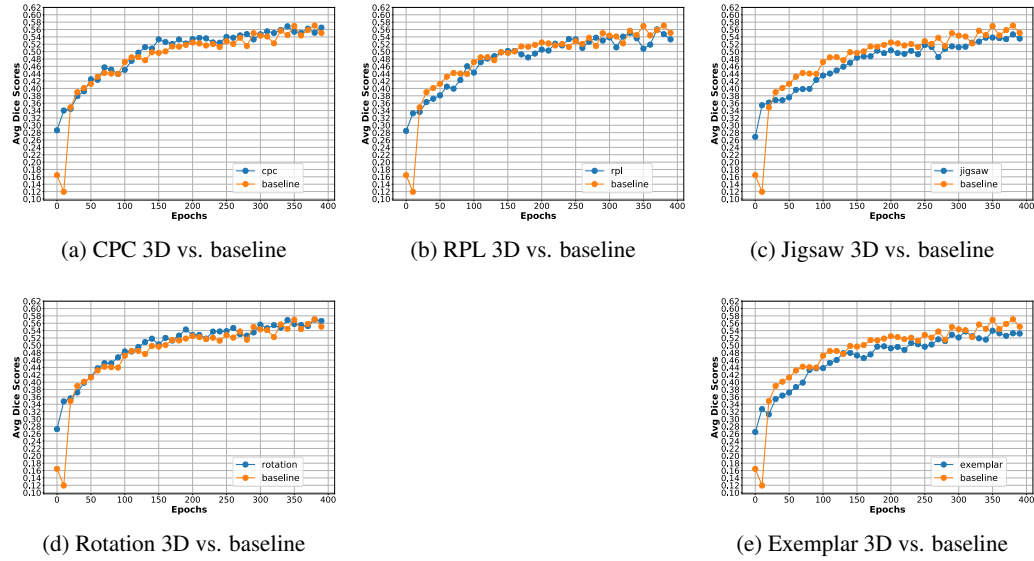


Figure 8: Pancreas segmentation: Detailed speed of convergence results per method (blue) vs. the supervised baseline (orange). This benefit of our methods helps achieve high results using only few epochs



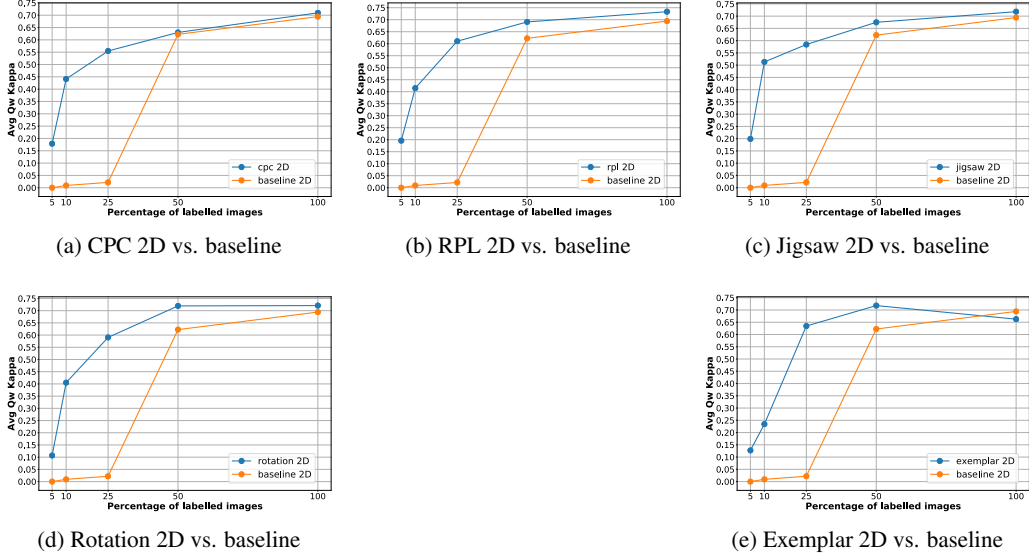


Figure 9: Retinopathy detection: Detailed data-efficiency results per method (blue) vs. the supervised baseline (orange). Our methods consistently outperform the baseline in low-data cases

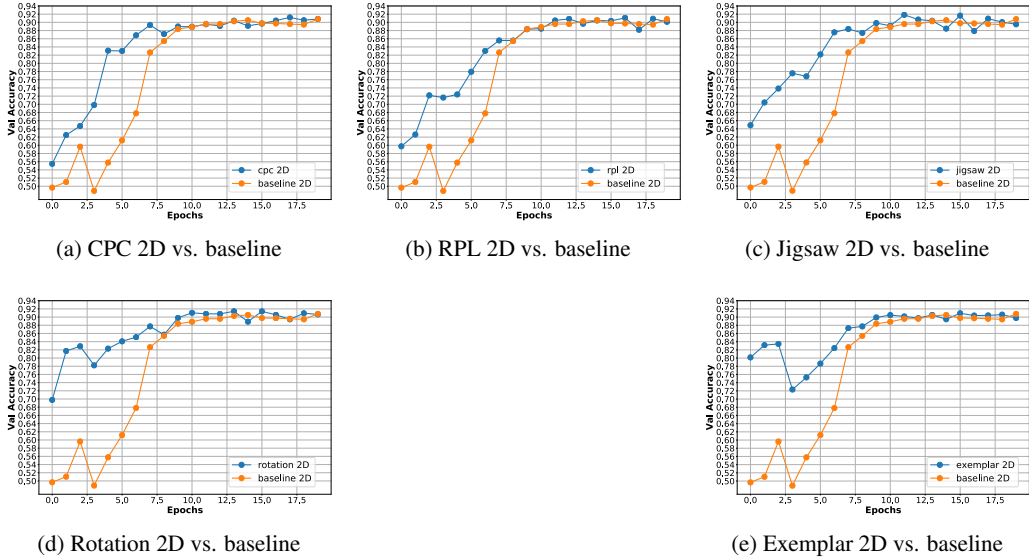


Figure 10: Retinopathy detection: Detailed speed of convergence results per method (blue) vs. the supervised baseline (orange). This benefit of our methods helps achieve high results using only few epochs