Transfer learning from natural image to medical image has been established as one of the most practical paradigms in deep learning for medical image analysis. To fit this paradigm, however, 3D imaging tasks in the most prominent imaging modalities (e.g., CT and MRI) have to be reformulated and solved in 2D, losing rich 3D anatomical information, thereby inevitably compromising its performance. To overcome this limitation, we have built a set of models, called Generic Autodidactic Models, nicknamed Models Genesis, because they are created ex nihilo (with no manual labeling), self-taught (learnt by self-supervision), and generic (served as source models for generating application-specific target models). Our extensive experiments demonstrate that our Models Genesis significantly outperform learning from scratch in all five target 3D applications covering both segmentation and classification. More importantly, learning a model from scratch simply in 3D may not necessarily yield performance better than transfer learning from ImageNet in 2D, but our Models Genesis consistently top any 2D/2.5D approaches including fine-tuning the models pre-trained from ImageNet as well as fine-tuning the 2D versions of our Models Genesis, confirming the importance of 3D anatomical information and significance of Models Genesis for 3D medical imaging. This performance is attributed to our unified self-supervised learning framework, built on a simple yet powerful observation: the sophisticated and recurrent anatomy in medical images can serve as strong yet free supervision signals for deep models to learn common anatomical representation automatically via self-supervision. As open science, all codes and pre-trained Models Genesis are available at https://github.com/MrGiovanni/ModelsGenesis.

© 2020 Elsevier B. V. All rights reserved.

1. Introduction

Transfer learning from natural image to medical image has become the de facto standard in deep learning for medical image analysis (Tajbakhsh et al., 2016; Shin et al., 2016), but given the marked differences between natural images and medical images, we hypothesize that transfer learning can yield more powerful (application-specific) target models from the source models built directly using medical images. To test this hypothesis, we have chosen chest imaging because the chest contains several critical organs, which are prone to a number of diseases that result in substantial morbidity and mortality, hence associated with significant health-care costs. In this research, we focus on Chest CT, because of its prominent role in diagnosing lung diseases, and our research community has accumulated several Chest CT image databases, for instance, LIDC-IDRI (Armato III et al., 2011) and NLST (NLST, 2011), containing a large number of Chest CT images. However, system-
Table 1: Pre-trained models with proxy tasks and target tasks. This paper uses transfer learning in a broader sense, where a source model is first trained to learn image presentation via full supervision or self-supervision by solving a problem, called proxy task (general or application-specific), on a source dataset with expert-provided or automatically-generated labels, and then this pre-trained source model is fine tuned (transferred) through full supervision to yield a target model to solve application-specific problems (target tasks) in the same or different datasets (target datasets). We refer transfer learning to same-domain transfer learning when the models are pre-trained and fine-tuned within the same domain (modality, organ, disease, or dataset), and to cross-domain when the models are pre-trained in one domain and fine-tuned for a different domain.

| Pre-trained model | Modality | Source dataset | Superv. / Annot. | Proxy task |
|-------------------|-----------|----------------|------------------|------------|
| Genesis Chest CT 2D | CT | LUNA 2016 (Setio et al., 2017) | Self / 0 | Image restoration on 2D Chest CT slices |
| Genesis Chest CT 3D | CT | LUNA 2016 (Setio et al., 2017) | Self / 0 | Image restoration on 3D Chest CT volumes |
| Genesis Chest X-ray | X-ray | ChestXray8 (Wang et al., 2017b) | Self / 0 | Image restoration on 2D Chest Radiographs |
| Models ImageNet | Natural | ImageNet (Deng et al., 2009) | Full / I4M images | Image classification on 2D ImageNet |
| Inflated 3D (3D) | Natural | Kinetics (Carreira and Zisserman, 2017) | Full / 240K videos | Action recognition on human action videos |
| NiftyNet | CT | Pancreas-CT & BTVC (Gibson et al., 2018a) | Full / 90 cases | Organ segmentation on abdominal CT |
| MedicalNet | CT, MRI | 3DSeg-8 (Chen et al., 2019b) | Full / 1,638 cases | Disease/organ segmentation on 8 datasets |

1 The first letter denotes the object of interest (“N” for lung nodule, “E” for pulmonary embolism, “L” for liver, etc); the second letter denotes the modality (“C” for CT, “M” for MRI, etc); the last letter denotes the task (“S” for classification, “O” for segmentation).

In summary, we make the following three contributions:

1. A collection of generic pre-trained 3D models, performing effectively across diseases, organs, and modalities.

2. A scalable self-supervised learning framework, offering encoder for classification and encoder-decoder for segmentation.

3. A set of self-supervised training schemes, learning robust representation from multiple perspectives.

In the remainder of this paper, we first in Sec. 2 introduce our self-supervised learning framework for training Models Genesis, covering our four proposed image transformations with their learning perspectives, and describing the four unique properties of our Models Genesis. Sec. 3 details the training process of Models Genesis and the five target tasks for evaluating Models Genesis, while Sec. 4 summarizes the five major observations from our extensive experiments, demonstrating that our Models Genesis can serve as a primary source of transfer learning for 3D medical imaging. In Sec. 5, we discuss various aspects of Models Genesis, including their capabilities for same-domain and cross-domain transfer learning and their impact on the creation of a medical ImageNet, followed by a thorough review.
Zongwei Zhou et al. / Medical Image Analysis (2020) 3

Fig. 1: [Better viewed on-line, in color, and zoomed in for details] Our self-supervised learning framework aims to learn general-purpose image representation by recovering the original sub-volumes of images from their transformed ones. We first crop arbitrarily-size sub-volume \( x_i \) at a random location from an unlabeled CT image. Each sub-volume \( x_i \) can undergo at most three out of four transformations: non-linear, local-shuffling, outer-cutout, and inner-cutout, resulting in a transformed sub-volume \( \tilde{x}_i \). It should be noted that outer-cutout and inner-cutout are considered mutually exclusive. Therefore, in addition to the four original individual transformations, this process yields eight more transformations, including one identity mapping (\( \phi \) meaning none of the four individual transformations is selected) and seven combined transformations. A Model Genesis, an encoder-decoder architecture with skip connections in between, is trained to learn a common image representation by restoring the original sub-volume \( x_i \) (as ground truth) from the transformed one \( \tilde{x}_i \) (as input), in which the reconstruction loss (MSE) is computed between the model prediction \( x'_i \) and ground truth \( x_i \). Once trained, the encoder alone can be fine-tuned for target classification tasks; while the encoder and decoder together can be fine-tuned for target segmentation tasks.

of existing supervised and self-supervised representation learning approaches in medical imaging in Sec. 6. Finally, Sec. 7 concludes and outlines future extensions of Models Genesis.

2. Models Genesis

The objective of Models Genesis is to learn a common image representation that is transferable and generalizable across diseases, organs, and modalities. Fig. 1 depicts our self-supervised learning framework, which enables training 3D models from scratch using unlabeled images, consisting of three steps: (1) cropping sub-volumes from patient CT images, (2) deforming the sub-volumes, and (3) training a model to restore the original sub-volume. In the following sections, we first introduce the denotations of our self-supervised learning framework and then detail each of the training schemes with its learning objectives and perspectives, followed by a summary of the four unique properties of our Models Genesis.

2.1. Image restoration proxy task

Given a raw dataset consisting of \( N \) patient volumes, theoretically we can crop infinite number of sub-volumes from the dataset. In practice, we randomly generate a subset \( X = \{ x_1, x_2, ..., x_n \} \), which includes \( n \) number of sub-volumes and then apply image transformation function to these sub-volumes, yielding

\[
\tilde{X} = f(X),
\]

where \( \tilde{X} = \{ \tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n \} \) and \( f(\cdot) \) denotes a transformation function. Subsequently, a Model Genesis, being an encoder-decoder network with skip connections in between, will learn to approximate the function \( g(\cdot) \) which aims to map the trans-
formed sub-volumes $\tilde{X}$ back to their original ones $X$, that is,
\begin{equation}
g(\tilde{X}) = X = f^{-1}(\tilde{X}).
\end{equation}

To avoid heavy weight dedicated towers for each proxy task and to maximize parameter sharing in Models Genesis, we consolidate four self-supervised schemes into a single image restoration task, enabling models to learn robust image representation by restoring from various sets of image transformations. Our proposed framework includes four transformations: (1) non-linear, (2) local-shuffling, (3) outer-cutout, and (4) inner-cutout. Each transformation is independently applied to a sub-volume with a predefined probability, while outer-cutout and inner-cutout are considered mutually exclusive. Consequently, each sub-volume can undergo at most three of the above transformations, resulting in twelve possible transformed sub-volume (see step 2 in Fig. 1). For clarity, we further define a training scheme as the process that (1) transforms sub-volumes using any of the aforementioned transformations, and (2) trains a model to restore the original sub-volumes from the transformed ones. For convenience, we refer to an individual training scheme as the scheme using one particular individual transformation.

### 2.2. Image transformations and learning perspectives

1) Learning appearance via non-linear transformation. We propose a novel self-supervised training scheme based on non-linear translation, with which the model learns to restore the intensity values of an input image transformed with a set of non-linear functions. The rationale is that the absolute intensity values (i.e., Hounsfield units) in CT scans or relative intensity values in other imaging modalities convey important information about the underlying structures and organs (Buzug, 2011; Forbes, 2012). Hence, this training scheme enables the model to learn the appearance of the anatomic structures perceivable, we intentionally retain the non-linear intensity transformation function as monotonic, allowing pixels of different values to be assigned with new distinct values. To realize this idea, we use Bézier Curve (Mortenson, 1999), a smooth and monotonic transformation function, which is generated from two end points ($P_0$ and $P_3$) and two control points ($P_1$ and $P_2$), defined as:
\begin{equation}
B(t) = (1-t)^3P_0 + 3(1-t)^2tP_1 + 3(1-t)t^2P_2 + t^3P_3, \quad t \in [0, 1],
\end{equation}
where $t$ is a fractional value along the length of the line. In Fig. 2(a), we illustrate the original CT sub-volume (the left-most column) and its transformed ones based on different transformation functions. The corresponding transformation functions are shown in the top row. Notice that, when $P_0 = P_1$ and $P_2 = P_3$ the Bézier Curve is a linear function (shown in Columns 2, 5). Besides, we set $P_0 = (0, 0)$ and $P_3 = (1, 1)$ to get an increasing function (shown in Columns 2—4) and the opposite to get a decreasing function (shown in Columns 5—7). The control points are randomly generated for more variances (shown in Columns 3, 4, 6, 7). Before applying the transformation functions, in Genesis CT, we first clip the Hounsfield units values within the range of $[-1000, 1000]$ and then normalize each CT scan to $[0, 1]$, while in Genesis X-ray, we directly normalize each X-ray to $[0, 1]$ without intensity clipping.

2) Learning texture via local pixel shuffling. We propose local pixel shuffling to enrich local variations of a sub-volume without dramatically compromising its global structures, which encourages the model to learn the local boundaries and textures of objects. To be specific, for each input sub-volume, we randomly select 1,000 windows and then shuffle the pixels inside each window sequentially. Mathematically, let us consider a small window $\mathbf{W}$ with a size of $m \times n$. The local-shuffling acts on each window and can be formulated as
\begin{equation}
\tilde{\mathbf{W}} = \mathbf{P} \times \mathbf{W} \times \mathbf{P}',
\end{equation}

![Fig. 2: Illustration of the proposed image transformations and their learning perspectives.](image-url)
where \( \hat{W} \) is the transformed window, \( P \) and \( P' \) denote permutation metrics with the size of \( m \times m \) and \( n \times n \), respectively. Pre-multiplying \( W \) with \( P \) permutes the rows of the window \( W \), whereas post-multiplying \( W \) with \( P' \) results in the permutation of the columns of the window \( W \). The size of the local window determines the difficulty of proxy task. In practice, to preserve the global content of the image, we keep the window sizes smaller than the receptive field of the network, so that the network can learn much more robust image representation by “resetting” the original pixels positions. Note that our method is quite different from PatchShuffling (Kang et al., 2017), which is a regularization technique to avoid over-fitting. As shown in Fig. 2(b), local-shuffling within an extent keeps the objects perceivable, benefiting the deep neural network in learning invariant image representations by restoring the original sub-volumes. Unlike de-noising (Vincent et al., 2010) and in-painting (Pathak et al., 2016; Iizuka et al., 2017), our local-shuffling transformation does not intend to replace the pixel values with noise, which therefore preserves the identical global distributions to the original sub-volume.

3) Learning context via outer and inner cutouts. We devise outer-cutout as a new training scheme for self-supervised learning. To realize it, we generate an arbitrary number (≤ 10) of windows, with various sizes and aspect ratios, and superimpose them on top of each other, resulting in a single window of a complex shape. When applying this merged window to a sub-volume, we leave the sub-volume region inside the window exposed and mask its surrounding (i.e., outer-cutout) with a random number. Moreover, to prevent the task from being too difficult or even unsolvable, we limit the outer-cutout region to be less than 1/4 of the whole sub-volume. By restoring the outer-cutouts, the model will learn the global geometry and spatial layout of organs in medical images via extrapolating within each sub-volume. We have illustrated this process step by step in Fig. 2(c). The first and last columns denote the original sub-volumes and the final transformed sub-volumes, respectively.

Our self-supervised learning framework also utilizes inner-cutout as a training scheme, where we mask the inner window regions (i.e., inner-cutouts) and leave their surroundings exposed. By restoring the inner-cutouts, the model will learn local continuities of organs in medical images via interpolating within each sub-volume. Unlike Pathak et al. (2016), where in-painting is proposed as a proxy task by restoring only the central region of the image, we restore the entire sub-volume as the model output. Examples of inner-cutout are illustrated in Fig. 2(d). Similar to outer-cutout, the inner-cutout areas are also limited to be less than 1/4 of the whole sub-volume, in order to keep the task reasonably difficult.

2.3. Unique properties of Models Genesis

1) Autodidactic—requiring no manual labeling. Models Genesis are trained in a self-supervised manner with abundant unlabeled image datasets, demanding zero expert annotation effort. Consequently, Models Genesis are fundamentally different from traditional (fully) supervised transfer learning from ImageNet (Shin et al., 2016; Tajbakhsh et al., 2016), which offers modest benefits to 3D medical imaging applications as well as that from the existing pre-trained, full-supervised models including I3D (Carreira and Zisserman, 2017), NiftyNet (Gibson et al., 2018b), and MedicalNet (Chen et al., 2019b), which demand a volume of annotation effort to obtain the source models (statistics given in Table 1).

2) Robust—learning from multiple perspectives. Our combined approach trains Models Genesis from multiple perspectives (appearance, texture, context, etc.), leading to more robust models across all target tasks, as evidenced in Figure 3, where our combined approach is compared with our individual schemes. This eclectic approach, incorporating multiple tasks into a single image restoration task, empowers Models Genesis to learn more comprehensive representation.

3) Scalable—accommodating many training schemes. Consolidated into a single image restoration task, our novel self-supervised schemes share the same encoder and decoder during training. Had each task required its own decoder, due to limited memory on GPUs, our framework would have failed to accommodate a large number of self-supervised tasks. By unifying all tasks as a single image restoration task, any favorable transformation can be easily amended into our framework, overcoming the scalability issue associated with multi-task learning (Doersch and Zisserman, 2017), where the network heads are subject to the specific proxy tasks.

4) Generic—yielding diverse applications. Models Genesis, trained via a diverse set of self-supervised schemes, learn a general-purpose image representation that can be leveraged for a wide range of target tasks. Specifically, Models Genesis can be utilized to initialize the encoder for the target classification tasks and to initialize the encoder-decoder for the target segmentation tasks, while the existing self-supervised approaches are largely focused on providing encoder models only (Jing and Tian, 2019). As shown in Table 3, Models Genesis can be generalized across diseases (e.g., nodule, embolism, tumor), organs (e.g., lung, liver, brain), and modalities (e.g., CT and MRI), a generic behavior that sets us apart from all previous works in the literature where the representation is learned via a specific self-supervised task, and thus lack generality.

3. Experiments

3.1. Pre-training Models Genesis

Our Models Genesis are pre-trained from 623 Chest CT scans in LUNA 2016 (Setio et al., 2017) in a self-supervised manner. The reason that we decided not to use all 888 scans provided by this dataset was to avoid test-image leaks between proxy and target tasks, so that we can confidently use the rest of the images solely for testing Models Genesis as well as the target models, although Models Genesis are trained from only unlabeled images, involving no annotation shipped with the dataset. We first randomly crop sub-volumes, sized 64×64×32 pixels, from different locations. To extract more informative sub-volumes for training, we then intentionally exclude those which are empty (air) or contain full tissues. Our Models Genesis 2D are self-supervised pre-trained from LUNA 2016 (Setio et al., 2017) and ChestX-ray14 (Wang et al., 2017b) using 2D CT slices in an
Table 2: Genesis CT is pre-trained on only LUNA 2016 dataset (i.e., the source) and then fine-tuned for five distinct medical image applications (i.e., the targets). These target tasks are selected such that they show varying levels of semantic distance from the source, in terms of organs, diseases, and modalities, allowing us to investigate the transferability of the pre-trained weights of Genesis CT with respect to the domain distance. The cells checked by √ denote the properties that are different between the source and target datasets.

| Task      | Disease | Organ | Dataset | Modality |
|-----------|---------|-------|---------|----------|
| NCC       |         |       |         |          |
| LCS       | √       |       |         |          |
| BMS       | √       |       |         | x        |

axial view and X-ray images, respectively. For all proxy tasks and target tasks, the raw image intensities were normalized to the [0, 1] range before training. We use the mean square error (MSE) between input and output images as objective function for the proxy task of image restoration. As suggested by Pathak et al. (2016), the MSE loss is sufficient for representation learning, although the restored images may be blurry.

3.2. Fine-tuning Models Genesis

The pre-trained Models Genesis can be adapted to new imaging tasks through transfer learning or fine-tuning. There are three major transfer learning scenarios: (1) employing the encoder as a fixed feature extractor for a new dataset and following up with a linear classifier (e.g., Linear SVM or Softmax classifier), (2) taking the pre-trained encoder and appending a sequence of fully-connected (fc) layers for target classification tasks, and (3) taking the pre-trained encoder and decoder and replacing the last layer with a 1×1×1 convolutional layer for target segmentation tasks. For scenarios (2) and (3), it is possible to fine-tune all the layers of the model or to keep some of the earlier layers fixed, only fine-tuning some higher-level portion of the model. We have evaluated the performance of our self-supervised representation for transfer learning by fine-tuning all layers in the network. In the following, we examine Models Genesis on five distinct medical applications, covering classification and segmentation tasks in CT and MRI images with varying levels of semantic distance from the source (Chest CT) to the targets in terms of organs, diseases, and modalities (see Table 2) for investigating the transferability of Models Genesis.

1) Lung nodule false positive reduction (NCC): The dataset is provided by LUNA 2016 (Setio et al., 2017) and consists of 888 low-dose lung CTs with slice thickness less than 2.5mm. Patients are randomly assigned into a training set (445 cases), a validation set (178 cases), and a test set (265 cases). The dataset offers the annotations for a set of 5,510,166 candidate locations for the false positive reduction task, wherein true positives are labeled as “1” and false positives are labeled as “0”. Following the prior works (Setio et al., 2016; Sun et al., 2017b), we evaluate performance via Area Under the Curve (AUC) score on classifying true positives and false positives.

2) Lung nodule segmentation (NCS): The dataset is provided by the Lung Image Database Consortium image collection (LIDC-IDRI) (Armato III et al., 2011) and consists of 1,018 cases collected by seven academic centers and eight medical imaging companies. The cases were split into training (510), validation (100), and test (408) sets. Each case is a 3D CT scan and the nodules have been marked as volumetric binary masks. We have re-sampled the volumes to 1-1-1 spacing and then extracted a 64×64×32 crop around each nodule. These 3D crops are used for model training and evaluation. As in prior works (Aresta et al., 2019; Tang et al., 2019; Zhou et al., 2018), we adopt Intersection over Union (IoU) and Dice coefficient scores to evaluate performance.

3) Pulmonary embolism false positive reduction (ECC): We utilize a database consisting of 121 computed tomography pulmonary angiography (CTPA) scans with a total of 326 emboli. The dataset is pre-processed as suggested in Zhou et al. (2017); Tajbakhsh et al. (2016, 2019b), divided at the patient-level into a training set with 434 true positive PE candidates and 3,406 false positive PE candidates, and a test set with 253 true positive PE candidates and 2,162 false positive PE candidates. To conduct a fair comparison with the prior works, we compute candidate-level AUC on classifying true positives and false positives.

4) Liver segmentation (LCS): The dataset is provided by MICCAI 2017 LiTS Challenge and consists of 130 labeled CT scans, which we split into training (100 patients), validation (15 patients), and test (15 patients) subsets. The ground truth segmentation provides two different labels: liver and lesion. For our experiments, we only consider liver as positive class and others as negative class and evaluate segmentation performance using Intersection over Union (IoU) and Dice coefficient scores.

5) Brain tumor segmentation (BMS): The dataset is provided by BraTS 2018 challenge (Menzle et al., 2015; Bakas et al., 2018) and consists of 285 patients (210 HGG and 75 LGG), each with four 3D MRI modalities (T1, T1c, T2, and Flair) rigidly aligned. We adopt 3-fold cross validation, in which two folds (190 patients) are for training and one fold (95 patients) for test. Annotations include background (label 0) and three tumor subregions: GD-enhancing tumor (label 4), the peritumoral edema (label 2), and the necrotic and non-enhancing tumor core (label 1). We consider those with label 0 as negatives and others as positives and evaluate segmentation performance using Intersection over Union (IoU) and Dice coefficient scores.

3.3. Baselines and implementation

For a thorough comparison, we used three different techniques to randomly initialize the weights of models: (1) a basic random initialization method based on Gaussian distributions, (2) a method commonly known as Xavier, which was suggested in Glorot and Bengio (2010), and (3) a revised version of Xavier called MSRA, which was suggested in He et al. (2015). They are implemented as uniform, glorot_uniform, and he_uniform, respectively, following the Initializers1 in Keras. Considering that most of the self-supervised learning methods are initially proposed and implemented in 2D, we have

1Initializers: faroit.com/keras-docs/1.2.2/initializations
extended two most representative ones (Vincent et al., 2010; Chen et al., 2019) into their 3D versions for a fair comparison. In addition, we have examined publicly available pretrained models for 3D transfer learning in medical imaging, including NiftyNet\(^2\) (Gibson et al., 2018b), MedicalNet\(^3\) (Chen et al., 2019b), and, the most influential 2D weights initialization, Models ImageNet. We also fine-tune 3D\(^4\) (Carreira and Zisserman, 2017) in our five target tasks because it has been shown to successfully initialize 3D models for lung nodule detection in Ardila et al. (2019).

3D U-Net architecture\(^5\) is used in 3D applications; U-Net architecture\(^6\) is used in 2D applications. Batch normalization (Ioffe and Szegedy, 2015) is utilized in all 3D/2D deep models. For proxy tasks, SGD method (Zhang, 2004) with an initial learning rate of 1e0 is used for optimization. We use ReduceLROnPlateau to schedule learning rate, in which if no improvement is seen in the validation set for a certain number of epochs, the learning rate is reduced. For target tasks, Adam method (Kinga and Adam, 2015) with a learning rate of 1e − 3 is used for optimization, where \(\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e − 8\). We use early-stop mechanism on the validation set to avoid over-fitting. Simple yet heavy 3D data augmentation techniques are employed in all five target tasks, including random flipping, transposing, rotating, and adding Gaussian noise. We run each method ten times on all of the target tasks and report the average, standard deviation, and further present statistical analysis based on an independent two-sample t-test.

4. Results

In this section, we begin with an ablation study to compare the combined approach with each individual scheme, concluding that the combined approach tends to achieve more robust results and consistently exceeds any other training schemes. We then take our pre-trained model from the combined approach and present results on five 3D medical applications, comparing them against the state-of-the-art approaches found in recent supervised and self-supervised learning literature.

4.1. The combined learning scheme exceeds each individual

We have devised four individual training schemes by applying each of the transformations (i.e., non-linear, local-shuffling, outer-cutout, and inner-cutout) individually to a sub-volume and training the model to restore the original one. We compare each of these training schemes with identical-mapping, which does not involve any image transformation. In three out of the five target tasks, as shown in Fig. 3, the model pre-trained by identical-mapping scheme does not perform as well as random initialization. This undesired representation obtained via identical-mapping suggests that without any image transformation, the model would not benefit much from the proxy image restoration task. On the contrary, nearly all of the individual schemes offer higher target task performances than identical-mapping, demonstrating the significance of the four devised image transformations in learning image representation.

Although each of the individual schemes has established the capability in learning image representation, its empirical performance varies from task to task. That being said, given a target task, there is no clear winner among the four individual schemes that can always guarantee the highest performance. As a result, we have further devised a combined scheme, which applies transformations to a sub-volume with a predefined probability for each transformation and trains a model to restore the original one. To demonstrate the importance of combining these image transformations together, we examine the combined training scheme against each of the individual ones. Fig. 3 shows that the combined scheme consistently exceeds any other individual schemes in all five target tasks. We have found that the combination of different transformations is advantageous because, as discussed, we cannot rely on one single training scheme to achieve the most robust and compelling results across multiple target tasks. It is our novel representation learning framework based on image restoration that allows integrating various training schemes into a single training scheme. Our qualitative assessment of image restoration quality, provided in Appendix Fig. A.9, further indicates that the combined scheme is superior over all four individual schemes in restoring the images that have been undergone multiple transformations. In summary, our combined scheme pre-trains a model from multiple perspectives (appearance, texture, context, etc.), empowering models to learn a more comprehensive representation, thereby leading to more robust target models. Based on the above ablation studies, in the following sections, we refer the models pre-trained by the combined scheme to Models Genesis and, in particular, refer the model pre-trained on LUNA 2016 dataset to Genesis Chest CT.

4.2. Models Genesis outperform learning from scratch

Transfer learning accelerates training and boosts performance, only if the image representation learned from the original (proxy) task is general and transferable to target tasks. Fine-tuning models trained on ImageNet has been a great success story in 2D (Tajbakhsh et al., 2016; Shin et al., 2016), but for 3D representation learning, there is no such a massive labeled dataset like ImageNet. As a result, it is still common practice to train 3D model from scratch in 3D medical imaging. Therefore, to establish the 3D baselines, we have trained 3D models with three representative random initialization methods, including naïve uniform initialization, Xavier/Glorot initialization proposed by Glorot and Bengio (2010), and He normal (MSRA) initialization proposed by He et al. (2015). In three out of the five 3D medical applications, the results reveal no significant difference among these random initialization methods, as presented in Fig. 4. Although randomly initializing weights can vary by the behaviors on different applications, He normal (MSRA), in which the weights are initialized with a specific ReLU-aware initialization, generally works the most reliably among all five target tasks. On the other hand, initialization

\(^2\)NiftyNet Model Zoo: github.com/NifTK/NiftyNetModelZoo
\(^3\)MedicalNet: github.com/Tencent/MedicalNet
\(^4\)3D: github.com/deepmind/kinetics-3d
\(^5\)3D U-Net: github.com/ellisdg/3DUnetCNN
\(^6\)Segmentation Models: github.com/qubvel/segmentation_models
with our pre-trained Genesis Chest CT achieves improvements over all three random initialization methods by a large margin. Our statistical analysis shows that the performance gain is significant for all the target tasks under study. This suggests that, owing to the representation learning scheme, our initial weights provide a better starting point than the ones generated under particular statistical distributions, while being over 13% faster (see Fig. 5). This observation has also been widely obtained in 2D model initialization (Tajbakhsh et al., 2016; Shin et al., 2016; Rawat and Wang, 2017; Zhou et al., 2017; Voulodimos et al., 2018). The hypothesis underneath transfer learning is that transferring deep features across visual tasks can obtain significant for all the target tasks under study. This suggests that, owing to the representation learning scheme, our initial weights provide a better starting point than the ones generated under particular statistical distributions, while being over 13% faster (see Fig. 5). This observation has also been widely obtained in 2D model initialization (Tajbakhsh et al., 2016; Shin et al., 2016; Rawat and Wang, 2017; Zhou et al., 2017; Voulodimos et al., 2018). The hypothesis underneath transfer learning is that transferring deep features across visual tasks can achieve a remarkable improvement over training from scratch, especially in the following section.

4.3. Models Genesis surpass existing pre-trained 3D models

We have evaluated our Models Genesis with existing publicly available pre-trained 3D models on five distinct medical target tasks. As shown in Table 3, Genesis Chest CT noticeably contrasts with any other existing 3D models, which have been pre-trained by full supervision. Note that, in the liver segmentation task (LCS), Genesis Chest CT is slightly outperformed by MedicalNet because of the benefit that MedicalNet gained from its (fully) supervised pre-training on the LiTS dataset directly. Further statistical tests reveal that Genesis Chest CT still yields comparable performance with MedicalNet at \( p = 0.05 \) level. For the rest four target tasks, Genesis Chest CT achieves superior performance against all its counterparts by a large margin, demonstrating the effectiveness and transferability of the learned features of Models Genesis, which are beneficial for both classification and segmentation tasks.

More importantly, although Genesis Chest CT is pre-trained on Chest CT only, it can generalize to different organs, diseases, datasets, and even modalities. As an example, the target task of pulmonary embolism false positive reduction is performed in Contrast-Enhanced CT scans that can appear differently from the proxy tasks in normal CT scans; yet, Genesis Chest CT achieves a remarkable improvement over training from scratch,
Fig. 5: Models Genesis enable better optimization than learning from scratch, evident by the learning curves for the target tasks of reducing false positives in detecting lung nodules (NCC) and pulmonary embolism (ECC) as well as segmenting lung nodule (LC5), liver (LC5), and brain tumor (BMS). We have plotted the validation performance averaged by ten trials for each application, in which accuracy and dice-coefficient scores are reported for classification and segmentation tasks, respectively. As seen, initializing with our pre-trained Models Genesis demonstrates benefits in the convergence speed.

| Pre-training  | Approach                               | Target tasks                |
|---------------|----------------------------------------|-----------------------------|
|               |                                        | NCC (%)                     |
|               |                                        | NCS (%)                     |
|               |                                        | ECC (%)                     |
|               |                                        | LCS (%)                     |
|               |                                        | BMS (%)                     |
| No            | Random with Uniform Init               | 94.74±1.97                  |
|               | Random with Xavier Init (Glorot and Bengio, 2010) | 75.48±0.43                  |
|               | Random with Xavier Init (Glorot and Bengio, 2010) | 74.05±1.97                  |
|               | Random with MSRA Init (He et al., 2015)   | 76.44±0.45                  |
|               | (Fully) supervised                     | I3D (Carreira and Zisserman, 2017) | 98.26±0.27                  |
|               |                                        | 71.58±0.55                  |
|               |                                        | 80.55±1.11                  |
|               |                                        | 70.65±4.26                  |
|               |                                        | 67.83±0.75                  |
|               | Self-supervised                        | NiftyNet (Gibson et al., 2018b) | 94.14±4.57                  |
|               |                                        | 52.98±2.05                  |
|               |                                        | 77.33±8.05                  |
|               |                                        | 83.23±1.05                  |
|               |                                        | 60.78±1.60                  |
|               | De-noising (revised in 3D) (Vincent et al., 2010) | 95.80±0.49                  |
|               |                                        | 75.68±0.32                  |
|               |                                        | 86.43±1.44                  |
|               |                                        | 85.52±0.58†                 |
|               |                                        | 66.09±1.35                  |
|               | Patch shuffling (revised in 3D) (Chen et al., 2019a) | 91.93±2.32                  |
|               |                                        | 75.74±0.51                  |
|               |                                        | 82.15±3.30                  |
|               |                                        | 82.82±2.35                  |
|               |                                        | 52.95±6.92                  |
|               | Rubiks Cube (revised) (Zhuang et al., 2019) | 96.24±1.27                  |
|               |                                        | 72.87±0.16                  |
|               |                                        | 80.49±4.64                  |
|               |                                        | 75.59±0.20                  |
|               |                                        | 62.75±1.93                  |
|               | Genesis Chest CT (ours)                | 98.34±0.44                  |
|               |                                        | 77.62±0.64                  |
|               |                                        | 87.20±2.87                  |
|               |                                        | 85.10±2.15                  |
|               |                                        | 67.96±1.29†††               |

1 The winner in LUNA (2016) holds an official score of 0.968 vs. 0.971 (ours)
2 Wu et al. (2018) holds a Dice of 74.05% vs. 75.86%±0.90% (ours)
3 Zhou et al. (2017) holds an AUC of 87.06% vs. 87.20%±2.87% (ours)
4 The winner in LiTS (2017) with post-processing holds a Dice of 96.60% vs. 93.19%±0.46% (ours without post-processing)
5 MRI Flair images are only utilized for segmenting brain tumors, so the results are not submitted to BraTS 2018.
6 Genesis Chest CT is slightly outperformed by MedicalNet in LCS because the latter has been (fully) supervised pre-trained on the LiTS dataset.

Increasing the AUC by 7 points. Moreover, Genesis Chest CT continues to yield a significant IoU gain in liver segmentation even though the proxy task and target task are significantly different in both, diseases affecting the organs (lung vs. liver) and the dataset itself (LUNA 2016 vs. LiTS 2017). We have further examined Genesis Chest CT and other existing pre-trained models using MRI Flair images, which represent the widest domain distance between the proxy and target tasks. As reported in Table 3 (BMS), Genesis Chest CT yields nearly a 5-point improvement in comparison with random initialization. The increased performance on the MRI imaging task is a particularly strong demonstration of the transfer learning capabilities of our Genesis Chest CT. To further investigate the behavior of Genesis Chest CT when encountering medical images from different modalities, we have provided extensive visualization in Appendix Fig. A.10, including example images from CT, X-ray, Ultrasound, and MRI modalities.

Considering the model footprint, our Models Genesis take the basic 3D U-Net as the backbone, carrying much fewer parameters than the existing open-source pre-trained 3D models. For example, we have adopted MedicalNet with resnet-101 as the backbone, which offers the highest performance based on Chen et al. (2019b) but comprises of 85.75M parameters; the pre-trained I3D (Carreira and Zisserman, 2017) contains 25.35M parameters in the encoder; the pre-trained NiftyNet uses Dense V-Networks (Gibson et al., 2019a) as backbone, comprising of only 2.6M parameters, but it does not perform as well as its counterparts in all five target tasks. Taken together, these results indicate that our Models Genesis, with only 16.32M parameters, surpass all existing pre-trained 3D models in terms of generalizability, transferability, and parameter efficiency.

4.4. Models Genesis reduce annotation efforts by at least 30%

While critics often stress the need for sufficiently large amounts of labeled data to train a deep model, transfer learning leverages the knowledge about medical images already learned by pre-trained models and therefore requires considerably fewer annotated data and training iterations than learning from scratch. We have simulated the scenarios of using a handful of labeled data, which allows investigating the power of our Models Genesis in transfer learning. Fig. 6 displays the results...
Fig. 6: Initializing with our Models Genesis, the annotation cost can be reduced by 30%, 50%, 57%, 84%, and 44% for target tasks NCC, NCS, ECC, LCS, and BMS, respectively. With decreasing amounts of labeled data, Models Genesis (red) retain a much higher performance on all five target tasks, whereas learning from scratch (grey) fails to generalize. Note that the horizontal red and gray lines refer to the performances that can eventually be achieved by Models Genesis and learning from scratch, respectively, when using the entire dataset.

Fig. 7: When solving problems in volumetric medical modalities, such as CT and MRI images, 3D volume-based approaches consistently offer superior performance than 2D slice-based approaches empowered by transfer learning. We conduct statistical analyses (circled in blue) between the highest performance achieved by 3D and 2D solutions. Training 3D models from scratch does not necessarily outperform their 2D counterparts (see NCC). However, training the same 3D models from Genesis Chest CT outperforms all their 2D counterparts, including fine-tuning Models ImageNet as well as fine-tuning our Genesis Chest X-ray and Genesis Chest CT 2D. It confirms the effectiveness of Genesis Chest CT in unlocking the power of 3D models. In addition, we have also provided statistical analyses between the highest and the second highest performances achieved by 2D models, finding that Models Genesis (2D) offer equivalent performances (n.s.) to Models ImageNet in four out of the five applications.

of training with a partial dataset, demonstrating that fine-tuning Models Genesis saturates quickly on the target tasks since it can achieve similar performance compared with the full dataset training. Specifically, with only 50%, 5%, 30%, 5%, and 30% of the labeled images, fine-tuning Models Genesis can approximate the performance achieved by learning from scratch using the entire dataset of NCC, NCS, ECC, LCS, and BMS, respectively. This shows that our Models Genesis can mitigate the lack of labeled images, resulting in a more annotation efficient deep learning in the end.

Furthermore, the performance gap between fine-tuning and learning from scratch is significant and steady over training models with each partial data point. For the lung nodule segmentation target task (LCS in Fig. 6), with 5% training data, Models Genesis can achieve the performance equivalent to learning from scratch using 10% training data. Based on this analysis, the cost of annotation in LCS can be reduced by half using Models Genesis compared with learning from scratch. For the pulmonary embolism false positive reduction target task (ECC), Fig. 6 suggests that with only 30% training samples, Models Genesis achieve performance equivalent to learning from scratch using 70% training samples. Therefore, nearly 57% of the labeling cost associated with the use of learning from scratch for ECC could be recovered with our Models Genesis. For the liver segmentation target task (LCS in Fig. 6), using 8% training data, Models Genesis equal the performance of learning from scratch using 50% training samples. Therefore, about 84% of the annotation cost associated with learning from scratch in LCS is recovered by initializing with Models Genesis. For the lung nodule segmentation target task (LCS in Fig. 6), with 5% training data, Models Genesis can achieve the performance equivalent to learning from scratch using 10% training data. Based on this analysis, the cost of annotation in LCS can be reduced by half using Models Genesis compared with learning from scratch. For the pulmonary embolism false positive reduction target task (ECC), Fig. 6 suggests that with only 30% training samples, Models Genesis achieve performance equivalent to learning from scratch using 70% training samples. Therefore, nearly 57% of the labeling cost associated with the use of learning from scratch for ECC could be recovered with our Models Genesis. For the liver segmentation target task (LCS in Fig. 6), using 8% training data, Models Genesis equal the performance of learning from scratch using 50% training samples. Therefore, about 84% of the annotation cost associated with learning from scratch in LCS is recovered by initializing.
Table 4: Our 3D approach, initialized by Models Genesis, significantly elevates the classification performance compared with 2.5D and 2D approaches in reducing lung nodule and pulmonary embolism false positives. The entries in bold highlight the best results achieved by different approaches. For the 2D slice-based approach, we extract input consisting of three adjacent axial views of the lung nodule or pulmonary embolism and some of their surroundings. For the 2.5D orthogonal approach, following the prior works presented in Roth et al. (2015); Tajbakhsh et al. (2015), each input is composed of an axial, coronal, and sagittal slice and centered at a lung nodule or pulmonary embolism candidate.

| Task | NCC | Random | ImageNet | Genesis |
|------|-----|--------|----------|---------|
| 2D slice-based input | 96.03±0.86 | 97.79±0.71 | 97.45±0.61 |
| 2.5D orthogonal input | 95.76±1.05 | 97.24±1.01 | 97.07±0.92 |
| 3D volume-based input | 96.03±1.82 | n/a | 98.34±0.44 |

| Task | NCC | Random | ImageNet | Genesis |
|------|-----|--------|----------|---------|
| 2D slice-based input | 60.33±8.61 | 62.57±8.04 | 62.84±8.78 |
| 2.5D orthogonal input | 71.27±4.64 | 78.61±3.73 | 78.58±3.67 |
| 3D volume-based input | 80.36±3.58 | n/a | 88.04±1.40 |

with Models Genesis. For the brain tumor segmentation target task (BMS) in Fig. 6, with less than 28% training data, Models Genesis achieve the performance equivalent to learning from scratch using 50% training data. Therefore, nearly 44% annotation efforts can be reduced using Models Genesis compared with learning from scratch. Overall, at least 30% annotation efforts have been reduced by Models Genesis, in comparison with learning a 3D model from scratch in five target tasks. With such annotation-efficient 3D transfer learning paradigm, computer-aided diagnosis of rare diseases, which are severely underrepresented owing to the difficulty of collecting a sizeable amount labeled data, could be eventually actualized.

4.5. Models Genesis consistently top any 2D/2.5D approaches

We have thus far presented the power of 3D models in processing volumetric data, in particular, with limited annotation. Besides adopting 3D models, another common strategy to handle limited data in volumetric medical imaging is to reformat 3D data into a 2D image representation followed by fine-tuning pre-trained Models ImageNet (Shin et al., 2016; Tajbakhsh et al., 2016). This approach increases the training examples by order of magnitude, but it sacrifices the 3D context. It is interesting to note how Genesis Chest CT compares with this standard in 2D. We have thus implemented two different methods to reformat 3D data into 2D input: the regular 2D representation obtained by extracting adjacent axial slices, and the 2.5D representation (Roth et al., 2014, 2015) composed of axial, coronal, and sagittal slices from volumetric data. Both of these 2D approaches seek to use 2D representation to emulate something three dimensional, in order to fit the paradigm of fine-tuning Models ImageNet. Fig. 7 exposes the comparison between 3D and 2D models on five 3D target tasks. Additionally, Table 4 compares 2D slice-based, 2.5D orthogonal, and 3D volume-based approaches on lung nodule and pulmonary embolism false positive reduction tasks. As evidenced by our statistical analyses, the 3D models trained from Genesis Chest CT achieve significantly higher average performance and lower standard deviation than 2D models fine-tuned from ImageNet using either 2D or 2.5D image representation. Nonetheless, the same conclusion does not apply to the models trained from scratch—3D scratch models are outperformed by 2D models in one out of the five target tasks (i.e., NCC in Fig. 7 and Table 4) and also exhibit an undeniably larger standard deviation. We attribute the mixed results of 3D scratch models to the larger number of model parameters and limited sample size in the target tasks, which together impede the full utilization of 3D context. In fact, the undesirable performance of the 3D scratch models highlights the effectiveness of Genesis Chest CT, which unlocks the power of 3D models for medical imaging. To summarize, we believe that 3D problems in medical imaging should be solved in 3D directly.

5. Discussions

5.1. Do we still need a medical ImageNet?

In computer vision, at the time this paper is written, no self-supervised learning method outperforms fine-tuning models pre-trained from ImageNet (Jing and Tian, 2019; Chen et al., 2019a; Kolesnikov et al., 2019; Zhou et al., 2019b; Hendrycks et al., 2019; Zhang et al., 2019; Caron et al., 2019). Therefore, it may seem surprising to observe from our results in Table 3 that (fully) supervised representation learning methods do not necessarily offer higher performances in some 3D target tasks than self-supervised representation learning methods do. We ascribe this phenomenon to the limited amount of supervision used in their pre-training (90 cases for NiftyNet (Gibson et al., 2018b) and 1,638 cases for MedicalNet (Chen et al., 2019b)) or the domain distance (from videos to CT/MRI for I3D (Carreira and Zisserman, 2017)). Evidenced by a prior study (Sun et al., 2017a) on ImageNet pre-training, large amount of supervision is required to foster a generic, comprehensive image representation. Back to 2009, when ImageNet had not been established, it was challenging to empower a deep model with generic image representation using a small or even medium size of labeled data, the same situation, we believe, that presents in 3D medical image analysis today. Therefore, despite the outstanding performance of Models Genesis, there is no doubt that a large, strongly annotated dataset for medical image analysis, like ImageNet (Deng et al., 2009) for computer vision, is still highly demanded. One of our goals for developing Models Genesis is to help create such a medical ImageNet, because based on a small set of expert annotations, models fine-tuned from Models Genesis will be able to help quickly generate initial rough annotations of unlabeled images for expert review, thus reducing the annotation efforts and accelerating the creation of a large, strongly annotated, medical ImageNet. In summary, Models Genesis are not designed to replace such a large, strongly annotated dataset for medical image analysis, like ImageNet for computer vision, but rather to help create one.

5.2. Same-domain or cross-domain transfer learning?

Same-domain transfer learning is always preferred whenever possible because a relatively smaller domain gap makes the learned image representation more beneficial for target tasks. Even the most recent self-supervised learning approaches in
medical imaging were solely evaluated within the same dataset, such as Chen et al. (2019a); Tajbakhsh et al. (2019a). Same-domain transfer learning strikes as a preferred choice in terms of performance; however, most of the existing medical datasets, with less than hundred cases, are usually too small for deep models to learn reliable image representation. Therefore, as our future work, we plan to combine the publicly available datasets from similar domains together to train modality-oriented models, including Genesis CT, Genesis MRI, Genesis X-ray, and Genesis Ultrasound, as well as organ-oriented models, including Genesis Brain, Genesis Lung, Genesis Heart, and Genesis Liver.

Cross-domain transfer learning in medical imaging is the Holy Grail. Retrieving a large number of unlabeled images from a PACS system requires an IRB approval, often a long process; the retrieved images must be de-identified; organizing the de-identified images in a way suitable for deep learning is tedious and laborious. Therefore, large quantities of unlabeled datasets may not be readily available to many target domains. Evidenced by our results in Table 3 (BMS), Models Genesis have a great potential for cross-domain transfer learning; particularly, our distortion-based approaches (such as non-linear and local-shuffling) take advantage of relative intensity values (in all modalities) to learn shapes and appearances of various organs. Therefore, as our future work, we will be focusing on methods that generalize well across domains.

6. Related Work

With the splendid success of deep neural networks, transfer learning (Pan and Yang, 2010; Weiss et al., 2016; Yosinski et al., 2014) has become integral to many applications, especially medical imaging (Greenspan et al., 2016; Litjens et al., 2017; Lu et al., 2017; Shen et al., 2017; Wang et al., 2017a; Zhou et al., 2017, 2019b). This immense popularity of transfer learning is attributed to the learned image representation offering convergence speedups and performance gains for most target tasks, in particular, with limited annotated data. In the following sections, we review the works related to supervised and self-supervised representation learning.

6.1. Supervised representation learning

ImageNet contains more than fourteen million images that have been manually annotated to indicate which objects are present in each image; and more than one million of the images have actually been annotated with the bounding boxes of the objects in the image. Pre-training a model on ImageNet and then fine-tuning it on different medical imaging tasks has seen the most practical adoption in the medical image analysis (Shin et al., 2016; Tajbakhsh et al., 2016). To classify the common thoracic diseases from ChestX-ray14 dataset, as evidenced in Irvin et al. (2019), nearly all the leading methods (Guan and Huang, 2018; Guendel et al., 2018; Ma et al., 2019; Tang et al., 2018) follow the paradigm of “fine-tuning Models ImageNet” by adopting different architectures, such as ResNet (He et al., 2016) and DenseNet (Huang et al., 2017), along with their pre-trained weights. Other representative medical applications include identifying skin cancer from dermatologist level photographs (Esteva et al., 2017), offering early detection of Alzheimer’s Disease (Ding et al., 2018), and performing effective detection of pulmonary embolism (Tajbakhsh et al., 2019b).

Despite the remarkable transferability of Models ImageNet, pre-trained 2D models offer little benefits towards 3D medical imaging tasks in the most prominent medical modalities (e.g., CT and MRI). To fit this paradigm, 3D imaging tasks have to be reformulated and solved in 2D or 2.5D (Roth et al., 2015, 2014; Tajbakhsh et al., 2015), thus losing rich 3D anatomical information and inevitably compromising the performance. Annotating 3D medical images at the similar scale with ImageNet requires a significant research effort and budget. It is currently not feasible to create annotated datasets comparable to this size for every 3D medical application. Consequently, for lung cancer risk malignancy estimation, Ardila et al. (2019) reported to incorporate 3D spatial information by using Inflated 3D (13D) (Carreira and Zisserman, 2017), trained from the Kinetics dataset, as the feature extractor. Evidenced by Table 3, it is not the most favorable choice owing to the large domain gap between the temporal video and medical volume. This limitation has led to the development of model zoo in NiftyNet (Gibson et al., 2018b). However, they were trained with small datasets for specific applications (e.g., brain parcellation and organ segmentation), and were never intended as source models for transfer learning. Our experimental results in Table 3 indicate that NiftyNet models offer limited benefits to the five target medical applications via transfer learning. More recently, Chen et al. (2019b) have pre-trained 3D residual network by jointly segmenting the objects annotated in a collection of eight medical datasets, resulting in MedicalNet for 3D transfer learning. In Table 3, we have examined the pre-trained MedicalNet on five target tasks in comparison with our Models Genesis. As reviewed, each and every aforementioned pre-trained model requires massive, high-quality annotated datasets. However, seldom do we have a perfectly-sized and systematically-labeled dataset to pre-train a deep model in medical imaging, where both data and annotations are expensive to acquire. We overcome the above limitation via self-supervised learning, which allows models to learn image representation from abundant unlabeled medical image data with zero human annotation effort.

6.2. Self-supervised representation learning

Aiming at learning image representation from unlabeled data, self-supervised learning research has recently experienced a surge in computer vision (Caron et al., 2018; Chen et al., 2019c; Doersch et al., 2015; Goyal et al., 2019; Jing and Tian, 2019; Mahendran et al., 2018; Mundhenk et al., 2018; Noroozi et al., 2018; Noroozi and Favaro, 2016; Pathak et al., 2016; Sayed et al., 2018; Zhang et al., 2016, 2017), but it is a relatively new trend in modern medical imaging. The key challenge for self-supervised learning is identifying a suitable self-supervision task, i.e., generating input and output instance pairs from the data. Two of the preliminary studies include predicting the distance and 3D coordinates of two patches randomly sampled
from the same brain (Spitzer et al., 2018), identifying whether two scans belong to the same person, and predicting the level of vertebral bodies (Jamaludin et al., 2017). Nevertheless, these two works are incapable of learning representation from “self-supervision” because they demand auxiliary information and specialized data collection such as paired and registered images. By utilizing only the original pixel/voxel information shipped with data, several self-supervised learning schemes have been conducted for different medical applications: Ross et al. (2018) adopted colorization as proxy task, wherein color colonoscopy images are converted to gray-scale and then recovered using a conditional Generative Adversarial Network (GAN); Alex et al. (2017) pre-trained a stack of denoising auto-encoders, wherein the self-supervision was created by mapping the patches with the injected noise to the original patches; Chen et al. (2019a) designed image restoration as proxy task, wherein small regions were shuffled within images and then let models learn to restore the original ones; Zhuang et al. (2019) introduced a 3D representation learning proxy task by recovering the rearranged and rotated Rubik’s cube; and finally Tajbakhsh et al. (2019a) individualized self-supervised schemes for a set of target tasks. As seen, the previously discussed self-supervised learning schemes, both in computer vision and medical imaging, are developed individually for specific target tasks, therefore, the generalizability and robustness of the learned image representation have yet to be examined across multiple target tasks. To our knowledge, we are the first to investigate cross-domain self-supervised learning in medical imaging.

6.3. Our previous work

Zhou et al. (2019b) first presented generic autodidactic models for 3D medical imaging, which obtain common image representation that is transferable and generalizable across diseases, organs and modalities, overcoming the scalability issue associated with multiple tasks. This paper extends the preliminary version substantially with the following improvements.

1. We have introduced notations, formulas, and diagrams, as well as detailed methodology descriptions along with their learning objectives, for a succinct framework overview in Sec. 2.

2. We have extended the brain tumor segmentation experiment using MRI Flair images in Sec. 3.2, highlighting the transfer learning capabilities of Models Genesis from CT to MRI Flair domains.

3. We have conducted comprehensive ablation studies between the combined scheme and each of the individual learning schemes in Sec. 4.1, demonstrating that learning from multiple perspectives leads to a more robust target task performance.

4. We have investigated three different random initialization methods for 3D models in Sec. 4.2, suggesting that initializing with Models Genesis can offer much higher performances and faster convergences.

5. We have examined Models Genesis with the existing pre-trained 3D models on five distinct medical target tasks in Sec. 4.3, showing that with less parameters, Models Genesis surpass all publicly available 3D models in both generalizability and transferability.

6. We have provided experimental results on five target tasks using limited annotated data in Sec. 4.4, indicating that transfer learning from our Models Genesis can reduce annotation efforts by at least 30%.

7. We have investigated 3D sub-volume based approach compared with 2D approaches fine-tuning from Models ImageNet using 2D/2.5D representation, revealing the power of pre-trained 3D models in Sec. 4.5.

7. Conclusion

A key contribution of ours is a collection of generic source models, nicknamed Models Genesis, built directly from unlabelled 3D imaging data with our novel unified self-supervised method, for generating powerful application-specific target models through transfer learning. While the empirical results are strong, surpassing state-of-the-art performances in most of the applications, our goal is to extend our Models Genesis to modality-oriented models, such as Genesis MRI and Genesis Ultrasound, as well as organ-oriented models, such as Genesis Brain and Genesis Heart. We envision that Models Genesis may serve as a primary source of transfer learning for 3D medical imaging applications, in particular, with limited annotated data. To benefit the research community, we make the development of Models Genesis open science, releasing our codes and models to the public. Creating all Models Genesis, an ambitious undertaking, takes a village; therefore, we would like to invite researchers around the world to contribute to this effort, and hope that our collective efforts will lead to the Holy Grail of Models Genesis, all powerful across diseases, organs, and modalities.

Acknowledgments

This research has been supported partially by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant, and partially by the National Institutes of Health (NIH) under Award Number R01HL128785. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work has utilized the GPUs provided partially by the ASU Research Computing and partially by the Extreme Science and Engineering Discovery Environment (XSEDE) funded by the National Science Foundation (NSF) under grant number ACI-1548562. We thank F. Haghighi and M. R. Hosseinzadeh Taher for implementing the 3D version of patch-shuffling (Chen et al., 2019a) and working with Z. Guo in evaluating the performance of MedicalNet (Chen et al., 2019b); M. M. Rahman Siddiquee for examining NiftyNet (Gibson et al., 2018b) with our Models Genesis; P. Zhang for comparing two additional random initialization methods with our Models Genesis; N. Tajbakhsh for revising our conference paper;
References

Alex, V., Vaidhya, K., Thirunavukkarasu, S., Kesavadas, C., Krishnamurthi, G., 2017. Semisupervised learning using denoising autoencoders for brain lesion detection and segmentation. Journal of Medical Imaging 4, 041311.

Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al., 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine 25, 954–961.

Aresta, G., Jacobs, C., Araújo, T., Cunha, A., Ramos, L., van Ginneken, B., Campilho, A., 2019. iw-net: an automatic and minimalistic interactive lung nodule segmentation deep network. Scientific reports 9, 1–9.

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al., 2011. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics 38, 915–931.

Bakas, S., Reyes, M., Jakub, A., Bauer, S., Rempfler, M., Crimà, A., Shinohara, R.T., Berger, C., Hu, S.M., Rozycki, M., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brains challenge. arXiv preprint arXiv:1811.02629 .

Bicle, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al., 2019. The liver tumor segmentation benchmark (ltb). arXiv preprint arXiv:1901.04056 .

Buzug, T.M., 2011. Computed tomography, in: Springer Handbook of Medical Imaging 3, 158, 113–122.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2959–2968.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2959–2968.

Chen, S., Ma, K., Zheng, Y., 2019b. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 .

Chen, T., Zhai, X., Ritter, M., Lucic, M., Housby, N., 2019c. Self-supervised learning via auxiliary rotation loss, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12154–12163.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 248–255.

Ding, Y., Sohn, J.H., Kwaczyński, M.G., Trivedi, H., Harnish, R., Jenkins, N.W., Litiev, D., Copeland, T.P., Aboian, M.S., Mari Aparici, C., et al., 2018. A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. Radiology 290, 456–464.

Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1422–1430.

Doersch, C., Zisserman, A., 2017. Multi-task self-supervised visual learning, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2051–2060.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115.

Forbes, G.B., 2012. Human body composition: growth, aging, nutrition, and activity. Springer Science & Business Media.

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018a. Automatic multi-organ segmentation on abdominal ct with dense v-networks. IEEE transactions on medical imaging 37, 1822–1834.

Gibson, E., Li, W., Sudre, C., Pidon, L., Shakir, D.I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., et al., 2018b. Nifnynet: a deep-learning platform for medical imaging. Computer methods and programs in biomedicine 158, 113–122.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256.

Goyal, P., Mahajan, D., Gupta, A., Misra, I., 2019. Scaling and benchmarking self-supervised visual representation learning. arXiv preprint arXiv:1905.01225 .

Greenspan, H., van Ginneken, B., Sammers, R.M., 2016. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging 35, 1153–1159.

Guo, Q., Huang, Y., 2018. Multi-label chest x-ray image classification via category-wise residual attention learning. Pattern Recognition Letters .

Guendel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A., Comaniciu, D., 2018. Learning to recognize abnormalities in chest x-rays with location-aware dense networks, in: Iberoamerican Congress on Pattern Recognition, Springer, pp. 757–765.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Heinovirta, D., Mzaouka, M., Kadavath, S., Song, D., 2019. Using self-supervised learning can improve model robustness and uncertainty, in: Advances in Neural Information Processing Systems, pp. 15637–15648.

Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 3.

Hurst, R.T., Burke, R.F., Wissner, E., Roberts, A., Kendall, C.B., Lester, S.J., Somers, V., Goldman, M.E., Wu, Q., Khandheria, B., 2010. Incidence of subclinical atherosclerosis as a marker of cardiovascular risk in retired professional football players. The American journal of cardiology 105, 1107–1111.

Izuka, S., Simo-Serra, E., Ishikawa, H., 2017. Globally and locally consistent image completion. ACM Transactions on Graphics (ToG) 36, 107.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Bach, F., Blei, D. (Eds.), Proceedings of the 32nd International Conference on Machine Learning, PMLR, Lille, France. pp. 448–456. URL: http://proceedings.mlr.press/v37/ioffe15.html.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901.07031 .

Jamaludin, A., Kadir, T., Zisserman, A., 2017. Self-supervised learning for spinal mris, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 294–302.

Jing, L., Tian, Y., 2019. Self-supervised visual feature learning with deep neural networks: A survey. arXiv preprint arXiv:1902.06162.

Kang, G., Dong, X., Zheng, L., Yang, Y., 2017. Patchshuffle regularization. arXiv preprint arXiv:1707.07103 .

Kinga, D., Adam, J.B., 2015. Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR).

Kolesnikov, A., Zhai, X., Beyer, L., 2019. Revisiting self-supervised visual representation learning, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1920–1929.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

LiTS, 2017. Results of all submissions for liver segmentation. URL: https://competitions.codalab.org/competitions/17094#results.

Lu, L., Zheng, Y., Caniego, G., Yang, L., 2017. Deep learning and convolutional neural networks for medical image computing: precision medicine, high performance and large-scale datasets. Springer.

LUNA, 2016. Results of all submissions for nodule false positive reduction. URL: https://luna16.grand-challenge.org/results/.

Ma, Y., Zhou, Q., Chen, X., Lu, H., Zhao, Y., 2019. Multi-attention network for thoracic disease classification and localization, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing
Appendix A. Qualitative assessment of image restoration

As presented in Sec. 3.1, Genesis CT and Genesis X-ray are pre-trained on LUNA 2016 (Setio et al., 2017) and Chest X-ray8 (Wang et al., 2017b), respectively, using a series of self-supervised learning schemes with different image transformations; more examples of our four individual transformations (i.e., non-linear, local-shuffling, outer-cutout, and inner-cutout) are illustrated in Fig. A.8. To obtain a generic image representation, Models Genesis are trained to restore the original images from their transformed counterparts.

To assess the image restoration quality at the time of inference, we pass the transformed images to the models that have been trained with different self-supervised learning schemes, including four individual and one combined schemes. In our visualization, the input images have undergone four individual transformations as well as eight different combined transformations, including the identity mapping (i.e., no transformation). As shown in Fig. A.9, the combined scheme can restore the unseen image by handling a variety of transformation (framed in red), whereas the models trained with the individual scheme can only restore unseen images that have undergone the same transformation as they were trained on (framed in green). This qualitative observations is consistent with our experimental finding in Sec. 4.1, that the combined learning scheme achieves the more robust and superior results over the individual scheme in transfer learning.

In Fig. A.10, we have further provided a qualitative assessment of image restoration quality by Genesis CT and Genesis X-ray, across medical imaging modalities. In our visualization, the input images are selected from four different medical modalities, covering X-ray, CT, Ultrasound, and MRI. It is clear from the figure that even though the models are only trained on single image modality, they can largely maintain the texture and structures during restoration not only within the same modality but also across different modalities. These observations suggest that Models Genesis are of a great potential in transferring learned image presentation across diseases, organs, datasets, and modalities.
Fig. A.8: Illustration of the proposed four image transformations. For simplicity and clarity, we illustrate the transformation on a 2D CT slice, but our Genesis Chest CT is trained using 3D sub-volumes directly, transformed in a 3D manner; our 3D image transformations, with an exception of non-linear transformation, cannot be approximated in 2D. For ease of understanding, in (a) non-linear transformation, we have displayed an image undergoing different translating functions in Columns 2—7; in (b) local-shuffling, (c) outer-cutout, and (d) inner-cutout transformation, we have illustrated each of the processes step by step in Columns 2—6, where the first and last columns denote the original images and the final transformed images, respectively. In local-shuffling, a different window (b) is automatically generated and used in each step.
As discussed in Fig. 1, each original image $x_i$ can possibly undergo twelve different transformations. We test the models with all these possible twelve transformed images $\tilde{x}_i$. We specify types of the image transformation $f(\cdot)$ for each row and the training scheme $g(\cdot)$ for each column. First of all, it can be seen that the models trained with individual schemes can restore unseen images that have undergone the same transformation very well (framed in green), but fail to handle other transformations. Taking non-linear transformation $f_{NL}(\cdot)$ as an example, any individual training scheme besides non-linear transformation itself cannot invert the pixel intensity from transformed whitish to the original blackish. As expected, the model trained with the combined scheme successfully restores original images from various transformations (framed in red). Second, the model trained with the combined scheme shows it is superior to other models even if they are trained with and tested on the same transformation. For example, in the local-shuffling case $f_{LS}(\cdot)$, the image recovered from the local-shuffling pre-trained model $g_{LS}(\cdot)$ is noisy and lacks texture. However, the model trained with the combined scheme $g_{NL,LS,OC,IC}(\cdot)$ generates an image with more underlying structures, which demonstrates that learning with augmented tasks can even improve the performance on each of the individual tasks. Third, the model trained with the combined scheme significantly outperforms models trained with individual training schemes when restoring images that have undergone seven different combined transformations (Rows 6—12). As an example, the model trained with non-linear transformation $g_{NL}(\cdot)$ can only recover the intensity distribution in the transformed image undergone $f_{NL,IC}(\cdot)$ but leaves the inner cutouts unchanged. These observations suggest that the model trained with the proposed unified self-supervised learning framework can successfully learn general anatomical structures and yield promising transferability on different target tasks. The quality assessment of image restoration further confirms our experimental observation, provided in Sec. 4.1, that the combined learning scheme exceeds each individual in transfer learning.
Fig. A.10: The left and right panels visualize the qualitative assessment of image restoration quality by Genesis CT and Genesis X-ray, respectively, across medical imaging modalities. For testing, we use the pre-trained model to directly restore images from LUNA 2016 (CT) (Setio et al., 2017), ChestX-ray8 (X-ray) (Wang et al., 2017b), CIMT (Ultrasound) (Hurst et al., 2010; Zhou et al., 2019a), and BraTS (MRI) (Menze et al., 2015). Though the models are only trained on single image modality, they can largely maintain the texture and structures during restoration not only within the same modality (framed in red), but also across different modalities.