Please cite the paper as 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 International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 384-393. Springer, Cham, 2019.

This paper was awarded a Young Scientist Award at MICCAI 2019

Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis

Zongwei Zhou[1], Vatsal Sodha[1], Md Mahfuzur Rahman Siddiquee[1], Ruibin Feng[1], Nima Tajbakhsh[1], Michael B. Gotway[2], and Jianming Liang[1]

[1] Arizona State University, Scottsdale, AZ 85259 USA
{zongweiz,vasodha,mrahmans,rfeng12,ntajbakh,jianming.liang}@asu.edu
[2] Mayo Clinic, Scottsdale, AZ 85259 USA
Gotway.Michael@mayo.edu

Abstract. Transfer learning from *natural* image to *medical* image has established as one of the most practical paradigms in deep learning for medical image analysis. However, to fit this paradigm, 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 and inevitably compromising the 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 (learned 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 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 our 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 yet recurrent anatomy in medical images can serve as strong supervision signals for deep models to learn common anatomical representation automatically via self-supervision. As open science, all pre-trained Models Genesis are available at https://github.com/MrGiovanni/ModelsGenesis.

1 Introduction

Given the marked differences between *natural* images and *medical* images, we hypothesize that transfer learning can yield more powerful (application-specific) *target* models if the *source* models are built directly from medical images. To test this hypothesis, we have chosen chest imaging because the chest contains several

Table 1: Target tasks.

Code†	Object	Modality	Source	Description
NCC	Lung Nodule	CT	LUNA2016	Lung nodule false positive reduction
NCS	Lung Nodule	CT	LIDC-IDRI	Lung nodule segmentation
ECC	Pulmonary Embolism	CT	PE-CAD	Pulmonary embolism false positive reduction
LCS	Liver	CT	LiTS2017	Liver segmentation
DXC	Pulmonary Diseases	X-ray	ChestX-ray8	Eight pulmonary diseases classification
IUC	CIMT RoI	Ultrasound	UFL MCAEL	RoI, bulb, and background classification
BMS	Brain Tumor	MRI	BraTS2013	Brain tumor segmentation

† 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, "X" for X-ray, "U" for Ultrasound, etc); the last letter denotes the task ("C" for classification, "S" for segmentation).

critical organs, which are prone to a number of diseases that result in substantial morbidity and mortality and thus are 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[1] and NLST[2], containing a large number of Chest CT images. Therefore, we seek to answer the following question: *Can we utilize the large number of available Chest CT images without systematic annotation to train source models that can yield high-performance target models via transfer learning?*

To answer this question, we have developed a framework that trains generic, source models for 3D imaging. We call the models trained with our framework Generic Autodidactic Models, nicknamed Models Genesis, and refer to the model trained using Chest CT scans as Genesis Chest CT. As ablation studies, we have also trained a downgraded 2D version using 2D Chest CT slices, called Genesis Chest CT 2D. To demonstrate the effectiveness of Models Genesis in 2D applications, we have trained a 2D model based on ChestX-ray8[3], named as Genesis Chest X-ray.

Our extensive experiments detailed in Sec. 3 demonstrate that Models Genesis, including Genesis Chest CT, Genesis Chest CT 2D, and Genesis Chest X-ray, *significantly* outperform learning from scratch in all seven target tasks (see Table 1). As revealed in Table 4, learning from scratch simply in 3D may *not* necessarily yield performance better than fine-tuning state-of-the-art ImageNet models, but our Genesis Chest CT *consistently* top any 2D approaches including fine-tuning ImageNet models as well as fine-tuning our Genesis Chest X-ray and Genesis Chest CT 2D, confirming the importance of 3D anatomical information in Chest CT and significance of our self-supervised learning method in 3D medical image analysis.

This performance is attributable to the following key observation: medical imaging protocols typically focus on particular parts of the body for specific clinical purposes, resulting in images of similar anatomy. The sophisticated yet recurrent anatomy offers consistent patterns for self-supervised learning to discover common representation of a particular body part (the lungs in our case).

[1] https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI
[2] https://biometry.nci.nih.gov/cdas/nlst/
[3] https://nihcc.app.box.com/v/ChestXray-NIHCC

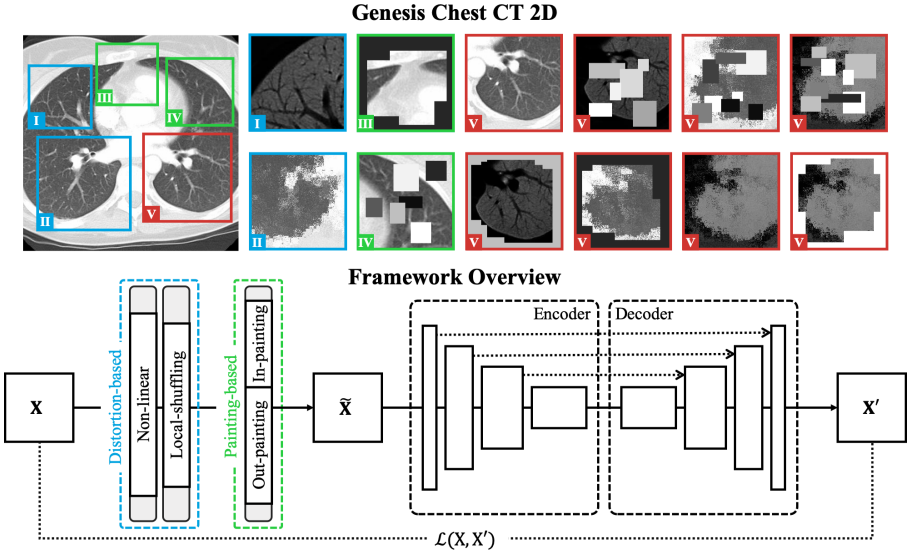


Fig. 1: Our unified self-supervised learning framework consolidates four novel transformations: I) non-linear, II) local-shuffling, III) out-painting, and IV) in-painting into a single image restoration task. Specifically, each arbitrarily-size patch X cropped at random location from an unlabeled image can undergo at most three of above transformations, resulting in a transformed patch $\tilde{X}$ (see I–V). Note that out-painting and in-painting are mutually exclusive. For simplicity and clarity, we illustrate our idea on a 2D CT slice, but our Genesis Chest CT is trained using 3D images directly. A Model Genesis, an encoder-decoder architecture, is trained to learn a common visual representation by restoring the original patch X (as ground truth) from the transformed one $\tilde{X}$ (as input), aiming to yield high-performance target models.

The fundamental idea behind our unified self-supervised learning method as illustrated in Fig. 1 is to recover anatomical patterns from images transformed via various ways in a unified framework.

2 Models Genesis

Models Genesis learn from scratch on unlabeled images, with an objective to yield a common visual representation that is generalizable and transferable across diseases, organs, and modalities. In Models Genesis, an encoder-decoder, as shown in Fig. 1, is trained using a series of self-supervised schemes. Once trained, the encoder alone can be fine-tuned for target classification tasks; while the encoder and decoder together can be for target segmentation tasks. For clarity, we formally define a *training scheme* as the process that transforms patches with any of the transformations, as illustrated in Fig. 1, and trains a model to restore the original patches from the transformed counterparts. In the following, we first explain each of our self-supervised learning schemes with its learning objectives and perspectives, followed by a summary of the four unique properties of our Models Genesis. Along the way, we also contrast Models Genesis with existing approaches to show our **innovations** and **novelties**.

- **Learning appearance via non-linear transformation.** Absolute or relative intensity values in medical images convey important information about the imaged structures and organs. For instance, the Hounsfield Units in CT scans correspond to specific substances of the human body. As such, intensity information can be used as a strong source of pixel-wise supervision. To preserve relative intensity information of anatomies during image transformation, we use Bézier Curve, a smooth and monotonous transformation function, which assigns every pixel a unique value, ensuring a one-to-one mapping. Restoring image patches distorted with non-linear transformation focuses Models Genesis on learning organ appearance (shape and intensity distribution). Fig. 1–I shows examples of the transformed images. Due to limited space, we provide the implementation details in Appendix[4] Sec. B.
- **Learning texture via local pixel shuffling.** Given an original patch, local pixel shuffling consists of sampling a random window from the patch followed by shuffling the order of contained pixels resulting in a transformed patch. The size of the local window determines the task difficulty, but we keep it smaller than the model's receptive field, and also small enough to prevent changing the global content of the image. Note that our method is quite different from PatchShuffling [5], which is a regularization technique to avoid over-fitting. To recover from local pixel shuffling, Models Genesis must memorize local boundaries and texture. Examples of local-shuffling are illustrated in Fig. 1–II. We include the underlying mathematics and implementation details in Appendix[4] Sec. C.
- **Learning context via out-painting and in-painting.** To realize the self-supervised learning via out-painting, we generate an arbitrary number of windows of various sizes and aspect ratios, and superimpose them on top of each other, resulting in a single window of a complex shape. We then assign a random value to all pixels outside the window while retaining the original intensities for the pixels within. As for in-painting, we retain the original intensities outside the window and replace the intensity values of the inner pixels with a constant value. Unlike [6], where in-painting is proposed as a proxy task by restoring only the patch central region, we restore the entire patch in the output. Out-painting compels Models Genesis to learn global geometry and spatial layout of organs via extrapolating, while in-painting requires Models Genesis to appreciate local continuities of organs via interpolating. Examples of out-painting and in-painting are shown in Fig. 1–III and Fig. 1–IV, respectively. More visualizations can be found in Appendix[4] Secs. D—E.

Models Genesis have the following four unique properties:

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 very different from traditional *supervised* transfer learning from ImageNet [7,9], which offers

[4] Appendix can be found in the full version at tinyurl.com/ModelsGenesisFullVersion

modest benefit to 3D medical imaging applications as well as that from the pre-trained models of NiftyNet[5], which is ineffective (see Sec. 3 and Appendix[4] Sec. I) due to the small datasets and specific applications (*e.g.,* brain parcellation and organ segmentation) these models are trained for.

2) Eclectic—learning from multiple perspectives. Our unified approach trains Models Genesis from multiple perspectives (appearance, texture, context, etc.), leading to more robust models across all target tasks, as evidenced in Table 3, where our unified 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—eliminating proxy-task-specific heads. 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 [2], where the network heads are subject to the specific proxy tasks.

4) Generic—yielding diverse applications. Models Genesis 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 [4]. As shown in Table 2, Models Genesis can be generalized across diseases (*e.g.,* nodule, embolism, tumor), organs (*e.g.,* lung, liver, brain), and modalities (*e.g.,* CT, X-ray, 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. Such specific schemes include predicting the distance and 3D coordinates of two patches randomly sampled from a same brain [8], identifying whether two scans belong to the same person, predicting the level of vertebral bodies [3], and finally the systematic study by Tajbakhsh *et al.* [10] where individualized self-supervised schemes are studied for a set of target tasks.

3 Experiments and Results

Experiment protocol. Our Genesis CT and Genesis X-ray are self-supervised pre-trained from 534 CT scans in LIDC-IDRI[1] and 77,074 X-rays in ChestX-ray8[3], respectively. The reason that we decided not to use all images in LIDC-IDRI and in ChestX-ray8 for training Models Genesis is to avoid test-image leaks between proxy and target tasks, so that we can confidently use the rest images solely for testing Models Genesis as well as the target models, although Models

[5] NiftyNet Model Zoo: https://github.com/NifTK/NiftyNetModelZoo

Table 2: Fine-tuning models from our Genesis Chest CT (3D) significantly outperforms learning from scratch in the five 3D target tasks ($p < 0.05$). The cells checked by ✗ denote the properties that are different between the proxy and target datasets. Our results show that our Genesis Chest CT generalizes across organs, diseases, datasets, and modalities. Footnotes show state-of-the-art performance for each target task.

Task	Metric	Disease	Organ	Dataset	Modality	Scratch (%)	Genesis (%)	p-value
NCC[1]	AUC					94.25±5.07	**98.20±0.51**	0.0180
NCS[2]	IoU					74.05±1.97	**77.62±0.64**	1.04e-4
ECC[3]	AUC	✗		✗		79.99±8.06	**88.04±1.40**	0.0058
LCS[4]	IoU	✗	✗	✗		74.60±4.57	**79.52±4.77**	0.0361
BMS[5]	IoU	✗	✗	✗	✗	90.16±0.41	**90.60±0.20**	0.0041

[1] LUNA winner holds an official score of 0.968 vs. 0.971 (ours)
[2] Wu *et al.* holds a Dice of 74.05% vs. 75.86%±0.90% (ours)
[3] Zhou *et al.* holds an AUC of 87.06% vs. 88.04%±1.40% (ours)
[4] LiTS winner w/ postprocessing (PP) holds a Dice of 96.60% vs. 91.13%±1.51% (ours w/o PP)
[5] BraTS winner w/ ensembling holds a Dice of 91.00% vs. 92.58%±0.30% (ours w/o ensembling)

Genesis are trained from *only* unlabeled images, involving *no* annotation shipped with the datasets. We evaluate Models Genesis in seven medical imaging applications including 3D and 2D image classification and segmentation tasks (codified as detailed in Table 1). For 3D applications in CT and MRI, we investigate the capability of both 2D slice-based solutions and 3D volume-based solutions; for 2D applications in X-ray and Ultrasound, we compare Models Genesis with random initialization and fine-tuning from ImageNet. 3D U-Net architecture[6] is used in five 3D applications; U-Net architecture with ResNet-18 encoder[7] is used in seven 2D applications. We utilize the L1-norm distance as the loss function in the image restoration tasks. Performances of target image classification and segmentation tasks are measured by the AUC (Area Under the Curve) and IoU (Intersection over Union), respectively, through at least 10 trials. We report the performance metrics with mean and standard deviation and further present statistical analysis based on the independent two-sample t-test.

Models Genesis outperform 3D models trained from scratch. We evaluate the effectiveness of Genesis Chest CT in five distinct 3D medical target tasks. These target tasks are selected such that they show varying levels of semantic distance to the proxy task, as shown in Table 2, allowing us to investigate the transferability of Genesis Chest CT with respect to the domain distance. Table 2 demonstrates that models fine-tuned from Genesis Chest CT consistently outperform their counterparts trained from scratch. Our statistical analysis show that the performance gain is significant for all the target tasks under study. Specifically, for NCC and NCS where the target and proxy tasks are in the same domain, initialization with Genesis Chest CT achieves 4 and 3 points increase in the AUC and IoU score, respectively, compared with training from scratch. For ECC, the target and proxy tasks are different in both the disease affecting the organ and the dataset itself; yet, Genesis Chest CT achieves a remarkable improvement over training from scratch, increasing the AUC by 8 points. Genesis Chest CT continues to yield significant IoU gain for LCS and BMS even though

[6] 3D U-Net Convolution Neural Network: https://github.com/ellisdg/3DUnetCNN
[7] Segmentation Models: https://github.com/qubvel/segmentation_models

Table 3: Comparison between our unified framework and each of the suggested self-supervised schemes on five 3D target tasks. The statistical analyses is conducted between the top-2 models in each column highlighted in red. While there is no clear winner, our unified framework is more robust across all target tasks, yielding either the best result or comparable performance to the best model ($p > 0.05$).

Approach	NCC (%)	NCS (%)	ECC (%)	LCS (%)	BMS (%)
Scratch	94.25±5.07	74.05±1.97	79.99±8.06	74.60±4.57	90.16±0.41
Distortion (ours)	96.46±1.03	77.08±0.68	**88.04±1.40**	79.08±4.26	**90.60±0.20**
Painting (ours)	**98.20±0.51**	77.02±0.58	87.18±2.72	78.62±4.05	90.46±0.21
Unified (ours)	97.90±0.57	**77.62±0.64**	87.20±2.87	**79.52±4.77**	90.59±0.21
p-value	0.0848	0.0520	0.2102	0.4249	0.4276

their domain distances with the proxy task are the widest. To our knowledge, we are the first to investigate cross-domain self-supervised learning in medical imaging. Given the fact that Genesis Chest CT is pre-trained on Check CT only, it is *remarkable* that our model can generalize to different diseases, organs, datasets, and even modalities.

Models Genesis consistently top any 2D approaches. A common technique to handle limited data in medical imaging is to reformat 3D data into a 2D image representation followed by fine-tuning pre-trained ImageNet models [7,9]. This approach increases the training examples by an order of magnitude, but it scarifies the 3D context. It is interesting to compare how Genesis Chest CT compares to this *de facto* standard in 2D. For this purpose, we adopt the trained 2D models from an ImageNet pre-trained model[7] for the tasks of NCC, NCS, and ECC. The 2D representation is obtained by extracting axial slices from volumetric datasets. Table 4 compares the results for 2D and 3D models. Note that the results for 3D models are identical to those reported in Table 2. As evidenced by our statistical analyses, the 3D models trained from Genesis Chest CT significantly outperform the 2D models trained from ImageNet, achieving higher average performance and lower standard deviation (see Table 4 and Appendix[4] Sec. H). However, the same conclusion does not apply to the models trained from scratch—3D scratch models outperform 2D scratch models in only two out of the three target tasks and also exhibit undesirably 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.

Models Genesis (2D) offer equivalent performances to supervised pre-trained models. To compare our self-supervised approaches with those supervised pre-training from ImageNet [1], we deliberately downgrade our Models Genesis to 2D versions: Genesis Chest CT 2D and Genesis Chest X-ray (2D) (see visualization of Genesis 2D in Appendix[4] Secs. F—G). The statistical analysis in Fig. 2 suggests that the downgraded Models Genesis 2D offer equivalent performance to state-of-the-art fine-tuning from ImageNet within modality, outperforming random initialization by a large margin, which is a significant achievement because ours comes at *zero* annotation cost. Meanwhile, the

Table 4: Comparison between 3D solutions and 2D slice-based solutions on three 3D target tasks. Training 3D models from scratch does not necessarily outperform the 2D counterparts (see NCC). However, training the same 3D models from Genesis Check CT outperforms ($p < 0.05$) all 2D solutions, demonstrating the effectiveness of Genesis Chest CT in unlocking the power of 3D models.

Task	2D (%)			3D (%)			p-value†
	Scratch	ImageNet	Genesis	Scratch	ImageNet	Genesis	
NCC	96.03±0.86	97.79±0.71	97.45±0.61	94.25±5.07	N/A	**98.20±0.51**	0.0213
NCS	70.48±1.07	72.39±0.77	72.20±0.67	74.05±1.97	N/A	**77.62±0.64**	<1e-8
ECC	71.27±4.64	78.61±3.73	78.58±3.67	79.99±8.06	N/A	**88.04±1.40**	5.50e-4

†These p-values are calculated between our Models Genesis vs. the fine-tuning from ImageNet, which always offers the best performance (highlighted in red) for all three tasks in 2D.

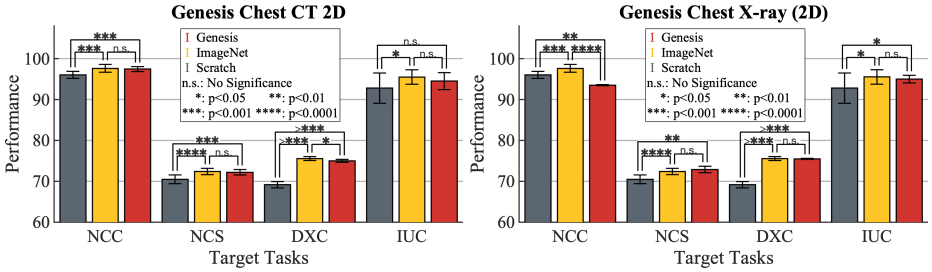


Fig. 2: Comparison of 2D solutions on four 2D target tasks. To investigate the same- and cross-domain transferability of Models Genesis, we have trained Genesis Chest CT 2D using 2D axial slices from LUNA dataset (left panel), and Genesis Chest X-ray (2D) trained using radiographs from ChestX-ray8 dataset (right panel). In same-domain target tasks (NCC and NCS in the left panel and DXC in the right panel), Models Genesis 2D outperform training from scratch and offer equivalent performance to fine-tuning from ImageNet. While in cross-domain target tasks (DXC and IUC in the left panel; NCS and IUC in the right panel), Models Genesis 2D also produce fairly robust performance.

downgraded Models Genesis 2D are fairly robust in cross-domain transfer learning, although they tend to underperform when domain distance is large, which suggests same-domain transfer learning should be preferred where possible in medical imaging. For 3D applications, we also examine the effectiveness of fine-tuning from NiftyNet[5], which is not designed for transfer learning but is the only available supervised pre-trained 3D model. Compared with training from scratch, fine-tuning NiftyNet suffers 3.37, 0.18, and 0.03 points decrease for NCS, LCS, and BMS tasks, respectively (detailed in Appendix[4] Sec. I), suggesting that strong supervision with limited annotated data cannot guarantee good transferability like ImageNet. Conversely, Models Genesis benefit from both large scale unlabeled datasets and dedicated proxy tasks which are essential for learning general-purpose visual representation.

4 Conclusion and Future Work

A key contribution of ours is a collection of *generic source* models, nicknamed Models Genesis, built directly from *unlabeled* 3D image data with our novel unified self-supervised method, for generating powerful application-specific *target*

models through transfer learning. While our empirical results are strong, surpassing state-of-the-art performances in most of the applications, an important future work 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. In fact, 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, and inviting researchers around the world to contribute to this effort. We hope that our collective efforts will lead to the Holy Grail of Models Genesis, effective 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 NIH under Award Number R01HL128785. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

1. Deng, J., *et al.*: ImageNet: A large-scale hierarchical image database. In: CVPR, 248–255 (2009) 7
2. Doersch, C., *et al.*: Multi-task self-supervised visual learning. In: ICCV, 2051–2060 (2017) 5
3. Jamaludin, A., *et al.*: Self-supervised learning for spinal MRIs. In: DLMIA, 294–302 (2017) 5
4. Jing, L., *et al.*: Self-supervised visual feature learning with deep neural networks: A survey. arXiv:1902.06162 (2019) 5
5. Kang, G., *et al.*: Patchshuffle regularization. arXiv:1707.07103 (2017) 4
6. Pathak, D., *et al.*: Context encoders: Feature learning by inpainting. In: CVPR, 2536–2544 (2016) 4
7. Shin, H.C., *et al.*: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. TMI, 35(5), 1285–1298 (2016) 4, 7
8. Spitzer, H., *et al.*: Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In: MICCAI, 663–671 (2018) 5
9. Tajbakhsh, N., *et al.*: Convolutional neural networks for medical image analysis: Full training or fine tuning? TMI, 35(5), 1299–1312 (2016) 4, 7
10. Tajbakhsh, N., *et al.*: Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. In: ISBI, 1251–1255 (2019) 5

Supplementary Material
Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis

Zongwei Zhou[1], Vatsal Sodha[1], Md Mahfuzur Rahman Siddiquee[1], Ruibin Feng[1], Nima Tajbakhsh[1], Michael B. Gotway[2], and Jianming Liang[1]

[1] Arizona State University, Scottsdale, AZ 85259 USA
{zongweiz,vasodha,mrahmans,rfeng12,ntajbakh,jianming.liang}@asu.edu
[2] Mayo Clinic, Scottsdale, AZ 85259 USA
Gotway.Michael@mayo.edu

Abstract. This document provides supplementary material for the paper entitled "Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis". The supplementary material is organized as follows. In Sec. A, we begin with a brief overview of Models Genesis. Secs. B—E describe at length the detailed implementation and illustration of four individual transformations. Secs. F—G contain a qualitative visualization on the pre-trained Genesis CT and Genesis X-ray for both same- and cross-domain image restoration. Secs. H—I present the transfer learning results of Models ImageNet, NiftyNet, and our Models Genesis in various target tasks.

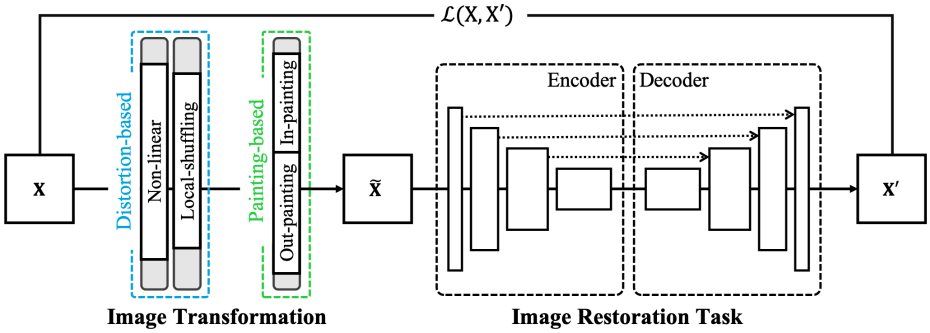


Fig. 3: Overview of our unified self-supervised learning framework. Given an image, we first extract patches X of arbitrary sizes from random locations and then apply the transformations on them as mentioned in Fig. 4. Models Genesis learns the visual representation by restoring the original patches X from the transformed ones $\tilde{X}$.

A Models Genesis

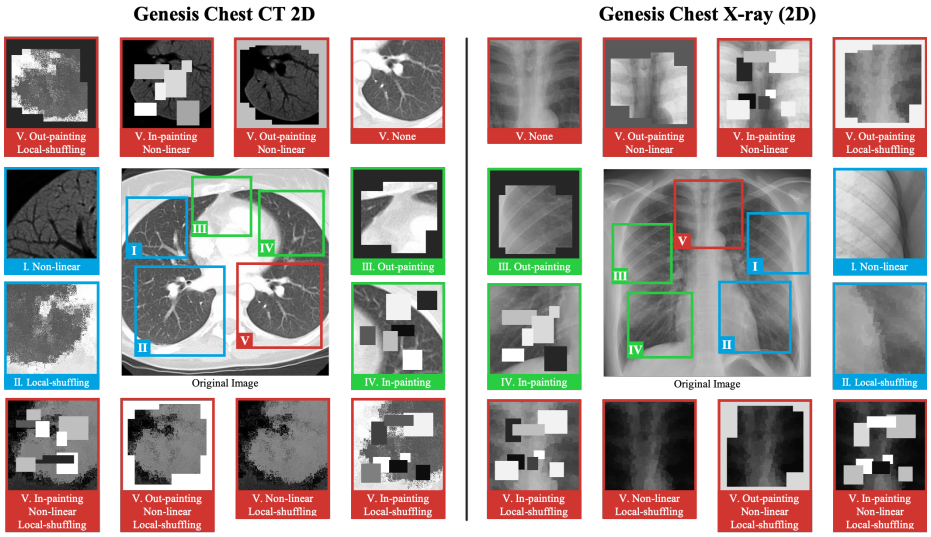


Fig. 4: [Better viewed on-line in color and zoomed in for details] Our novel unified self-supervised learning framework aims to learn general-purpose visual representation by recovering original image patches from their transformed ones. We have designed four individual transformations: I) non-linear (see Sec. B), II) local-shuffling (see Sec. C), III) out-painting (see Sec. D), and IV) in-painting (see Sec. E). We have provided examples of the transformed images for Genesis Chest CT (left) and Genesis Chest X-ray (right). For simplicity and clarity, we illustrate our idea on a 2D CT slice and a 2D X-ray image, but our Genesis Chest CT is trained using 3D Check CT images directly. Each transformation is independently applied to a patch with a predefined probability, while out-painting and in-painting are considered mutually exclusive. Therefore, in addition to the four original individual transformations, this process yields eight more transformations framed in red, including one identity mapping (*i.e.,* V: none, meaning none of the four individual transformations is selected) and seven combined transformations as indicated under each patch framed in red. For clarity, we further define a *training scheme* as the process that transforms patches with any of the twelve aforementioned transformations and trains a model to restore the original patches from the transformed ones. For convenience, we refer to an *individual training scheme* as the scheme using one particular individual transformation. Finally, our unified learning framework utilizes all possible transformations randomly with pre-defined probabilities and trains a model to restore the original patches from the ones undergone any possible transformations.

As shown in Fig. 3, our proposed self-supervised learning framework consists of two components: image transformation (illustrated in Fig. 4) and image restoration, where Models Genesis, taking an encoder-decoder architecture, are trained by restoring original patch X from transformed patch $\tilde{X}$, aiming to learn

common visual representation that is transferable and generalizable across diseases, organs and, modalities and thus yield high-performance target models. From this work, we have concluded:

1. 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 Models Genesis consistently top any 2D approaches including fine-tuning from ImageNet [11] as well as fine-tuning our 2D Models Genesis, confirming the importance of 3D anatomical information and significance of our Models Genesis for 3D medical imaging.
2. Despite the outstanding performance of Models Genesis, a large, strongly annotated dataset for medical image analysis like ImageNet [4] for computer vision is still highly demanded. In computer vision, at the time this paper is written, no self-supervised learning method outperforms fine-tuning models pre-trained from ImageNet [10,3,12]. One of our goals of developing Models Genesis is to help create such a large, strongly annotated dataset for medical image analysis, 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 helping create one.
3. Same-domain transfer learning is always preferred whenever possible. Same-domain transfer learning strikes as a preferred choice in terms of performance; therefore, as our future work, we continue training 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.
4. Cross-domain transfer learning 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 2 and Fig. 2, Models Genesis have a great potential for cross-domain transfer learning; particularly, distortion-based approaches 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 in cross-domain transfer learning. Building the Holy Grail of Models Genesis, effective across diseases, organs, and modalities, takes a village. As a result, we make the development of Models Genesis open science, inviting researchers around the world to join this effort. All pre-trained Models Genesis will be made public at https://github.com/MrGiovanni/ModelsGenesis.

B Non-linear Intensity Transformation Visualization

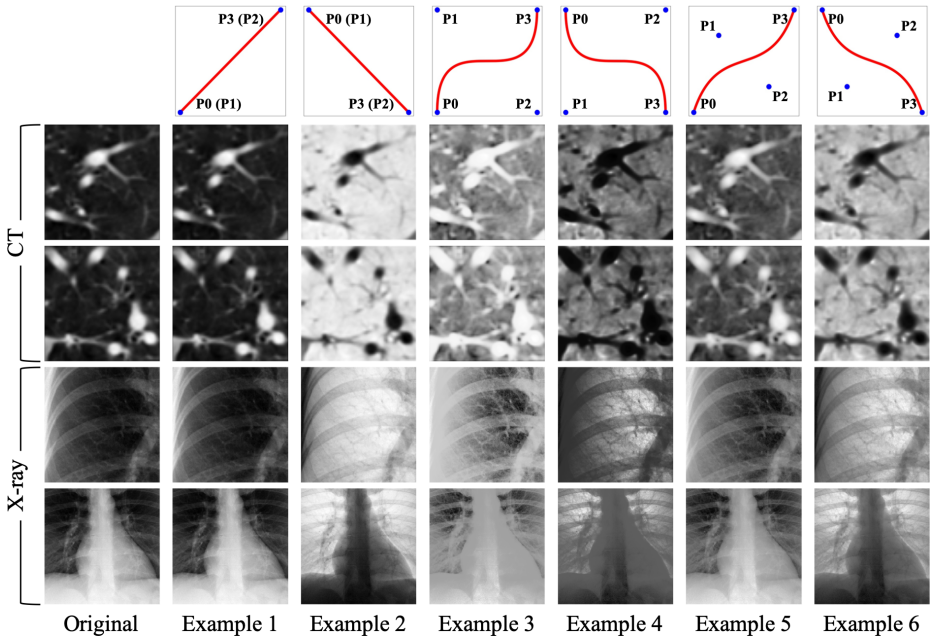


Fig. 5: We adopt non-linear intensity transformation as a new training scheme for self-supervised learning, which allows the model to learn the absolute or relative appearance of structures. Illustration of various non-linear intensity transformations (Examples 1—6) for CT (Rows 2—3) and X-ray (Rows 4—5) images is provided. We utilize four control points (P_0—P_3) in Eq. 1 to modify the shape of the transformation function (Row 1). Notice that, when $P_0 = P_1$ and $P_2 = P_3$ the transformation function is a linear function (shown in Examples 1—2). Besides, we set $P_0 = (0, 0)$ and $P_3 = (1, 1)$ to get an increasing function (shown in Examples 1, 3, and 5) and the opposite to get a decreasing function (shown in Examples 2, 4, and 6). The control points are randomly generated for more variances.

We propose a novel self-supervised training scheme based on non-linear translation, with which the model learns to restore the intensity values of the 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 [2,5]. Hence, this training scheme enables the model to learn the appearance of the anatomic structures present in the images. In order to keep the appearance of the anatomic structures perceivable, we keep the non-linear intensity transformation function *monotonic*, allowing pixels of different values to be assigned with new distinct values. To realize this idea, we use Bézier Curve [14], a smooth and monotonous 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:

$$B(t) = (1-t)^3 P_0 + 3(1-t)^2 t P_1 + 3(1-t) t^2 P_2 + t^3 P_3, \ t \in [0,1], \qquad (1)$$

where t is a fractional value along the length of the line. In Fig. 5, we illustrate the original patches (the left-most column) and the transformed patches of 2D CT and X-rays based on different transformation functions. The corresponding transformation functions are shown in the top row. In order to apply the transformation functions on CT images, we first clip the HU values to a range of $[-1000, 1000]$ and then normalize to $[0, 1]$ for each of the CT image slices. In contrast, the X-ray images are directly normalized to $[0, 1]$ without intensity clipping.

C Local Pixel Shuffling Visualization

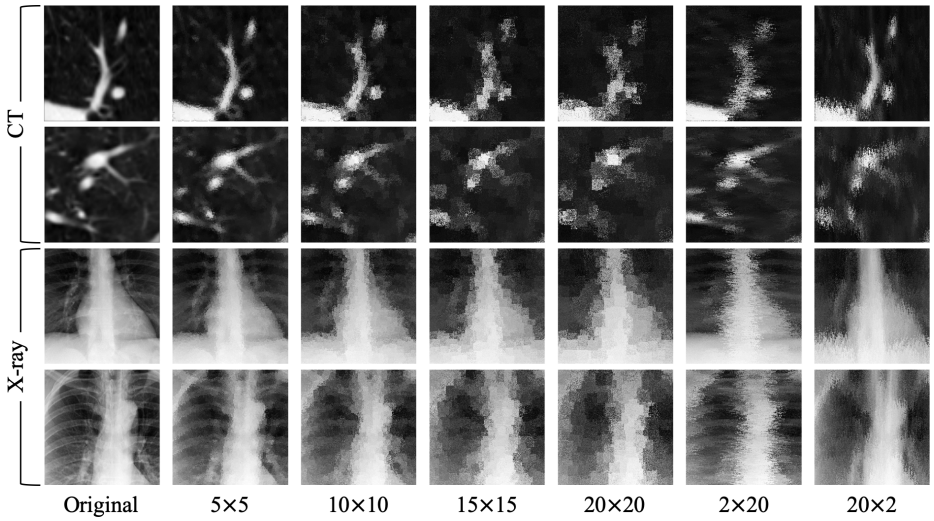


Fig. 6: We adopt local pixel shuffling as a new training scheme for self-supervised learning, which allows the model to learn the global geometry and spatial layout of organs as well as the local shape and texture of organs. Illustration of local pixel shuffling using multiple window sizes (Columns 2—7) applied on CT (Rows 1—2) and X-ray (Rows 3—4) images is provided. When 5×5 window is applied, the shapes are largely maintained; while the ribs are mostly invisible for window size equal to 20×20. Besides, various aspect ratios of windows also impose more local variances in different directions. Taking the restored X-ray patches in the last two columns as examples, a window size with $h \ll w$ (Column 6) distorts the boundary of the spine while preserving the overall presence of the ribs. On the other hand, when $h \gg w$ (Column 7), the ribs are hardly visible but the width of spine and heart is barely changed.

We propose local pixel shuffling to enrich local variations of a patch without dramatically compromising its global structures, which encourages the model to learn the shapes and boundaries of the objects as well as the relative layout of different parts of the objects. To be specific, for each input patch, we randomly select 1,000 windows from the patch and then shuffle the pixels inside each window sequentially. Mathematically, let us consider a small window $\mathbf{W}$ with the size of $m \times n$. The local-shuffling acts on each window and can be formulated as

$$\tilde{\mathbf{W}} = \mathbf{P} \times \mathbf{W} \times \mathbf{P}', \tag{2}$$

where $\tilde{\mathbf{W}}$ is the transformed window, $\mathbf{P}$ and $\mathbf{P}'$ denote permutation metrics with the size of $m \times m$ and $n \times n$, respectively. Pre-multiplying $\mathbf{W}$ with $\mathbf{P}$ permutes the rows of the window $\mathbf{W}$, whereas post-multiplying $\mathbf{W}$ with $\mathbf{P}'$ results in the permutation of the columns of the window $\mathbf{W}$. In practice, we keep the

window sizes smaller than the receptive field of the network, so that the network can learn a more robust visual representation by "resetting" the original pixel positions. To facilitate the understanding, we have explored the local-shuffling transformation of varying window sizes and illustrated them along with the original patches. The window sizes can control the degree of distortion. As shown in Fig. 6, local-shuffling within an extent keeps the objects perceivable, it will benefit the deep neural network in learning invariant visual representations by restoring the original patches. Unlike de-noising [16] and in-painting [15,9], our local-shuffling transformation does not intend to replace the pixel values with noise, which therefore preserves the identical global distributions to the original patch.

D Out-painting Visualization

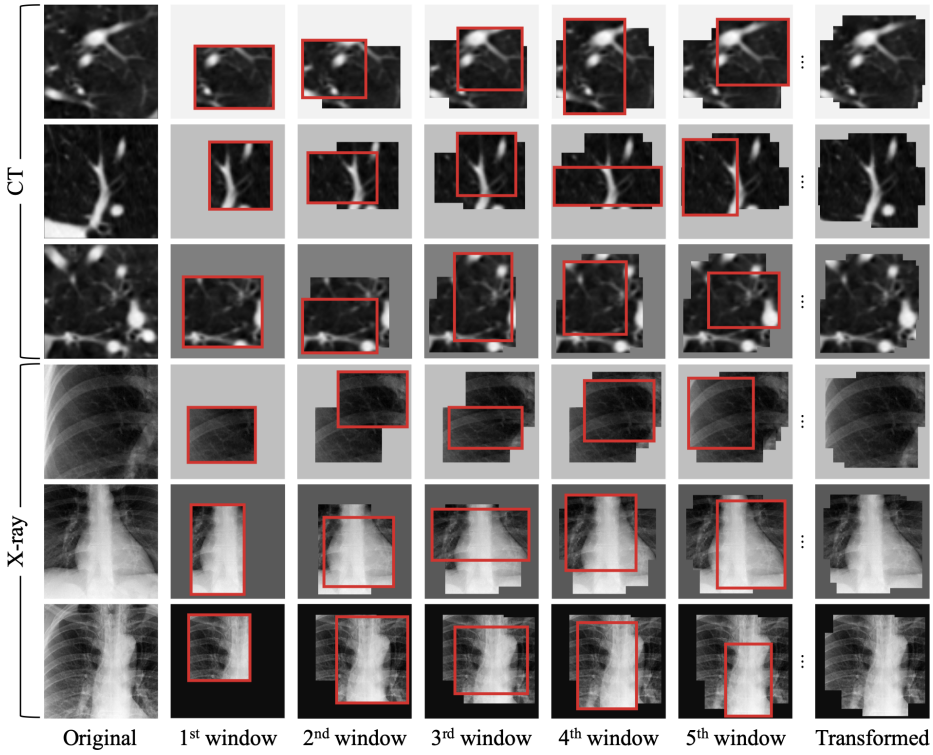


Fig. 7: We adopt out-painting as a new training scheme for self-supervised learning, which allows the model to learn the *global* geometry and spatial layout of organs. Illustration of the transformation for out-painting using various window sizes in CT (Rows 1—3) and X-ray (Rows 4—6) images is provided. The first and last columns denote the original patches and the final transformed patches, respectively. From Column 2 to Column 6, we generate a new window (red framed) and merge it with the existing ones. Moreover, to prevent the task to be too difficult or even unsolvable, we limit the masked surrounding region less than 1/4 of the whole patch.

We devise out-painting as a new training scheme for self-supervised learning, which allows the network to learn *global* geometry and spatial layout of organs in medical images by extrapolation. To realize it, we generate an arbitrary number ($\leqslant 10$) 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 patch, we leave the patch region inside the window exposed and mask its surrounding with a random number. We have illustrated this process step by step in Fig. 7.

E In-painting Visualization

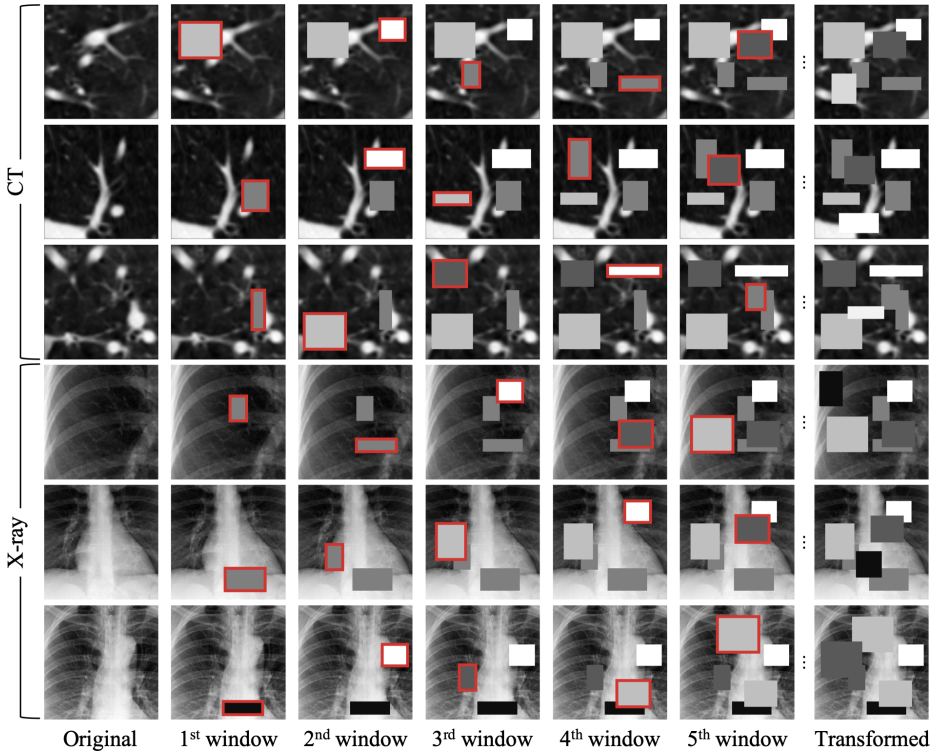


Fig. 8: We adopt in-painting as a new training scheme for self-supervised learning, which allows the model to learn *local* shape and texture of organs in medical images via interpolation. The final transformed patches (Column 7) are obtained by iteratively superimposing a window of random size and aspect ratio, filled with a random number, to the original patches (Column 1). Columns 2—6 illustrate this process step by step. Similar to out-painting, the masked areas are also limited to be less than 1/4 of the whole patch, in order to keep the task reasonably difficult.

F Genesis Chest CT

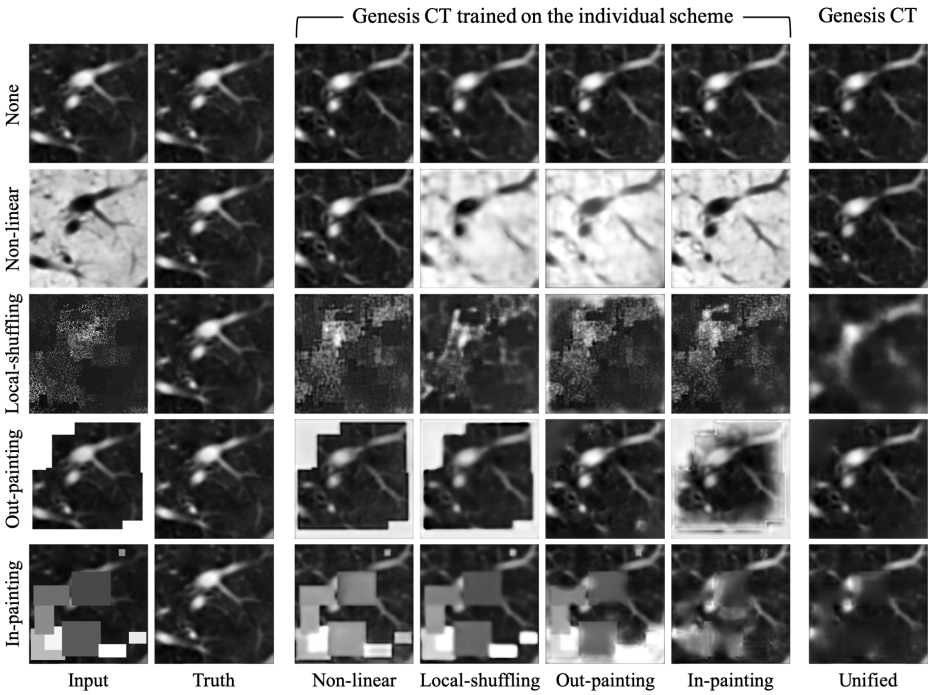


Fig. 9: Qualitative assessment of CT image restoration quality using Models Genesis trained with different training schemes, including the unified framework and four individual training schemes. LIDC-IDRI [1] is used for both training and testing. We test these models with transformed patches that have undergone four individual transformations as well as the identity mapping (*i.e.,* no transformation). First of all, it can be seen the models trained with single-transformation-based schemes fail to handle other transformations. Taking non-linear transformation (Row 2) 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 unified framework successfully restores original images from various transformations. Second, the model trained with the unified framework shows its superior to other models even if they are trained with and tested on the same transformation. For example, in local-shuffling case (Row 3), the patch recovered from the local-shuffling pre-trained model (Column 4) is noisy and lacks texture. However, the model trained with the unified framework (Column 7) generates a patch with more underlying structures, which demonstrates that learning with augmented tasks can even improve the performance on each individual tasks. 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.

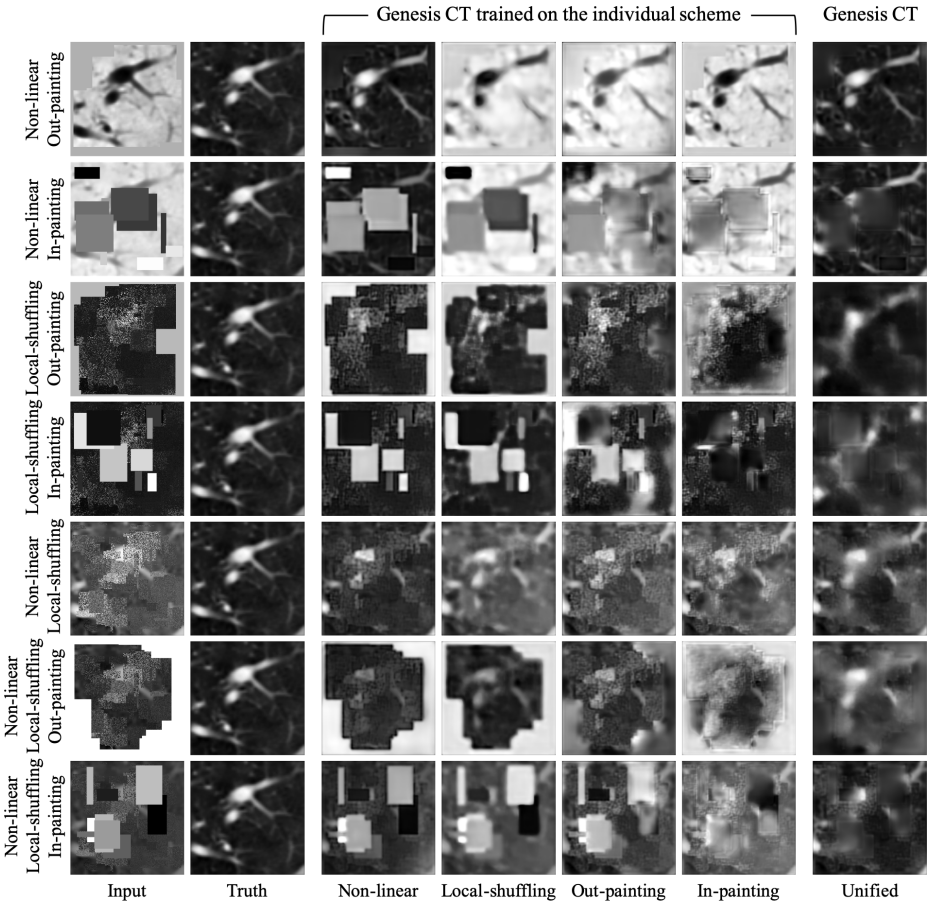


Fig. 10: Continued from Fig. 9. To further test our models, we show the restoration results on CT slices undergone seven different combined transformations. As expected, the model trained with our unified self-supervised framework significantly outperforms models trained with individual training schemes, further demonstrating the effectiveness of the proposed unified training framework as well as the pre-trained Models Genesis.

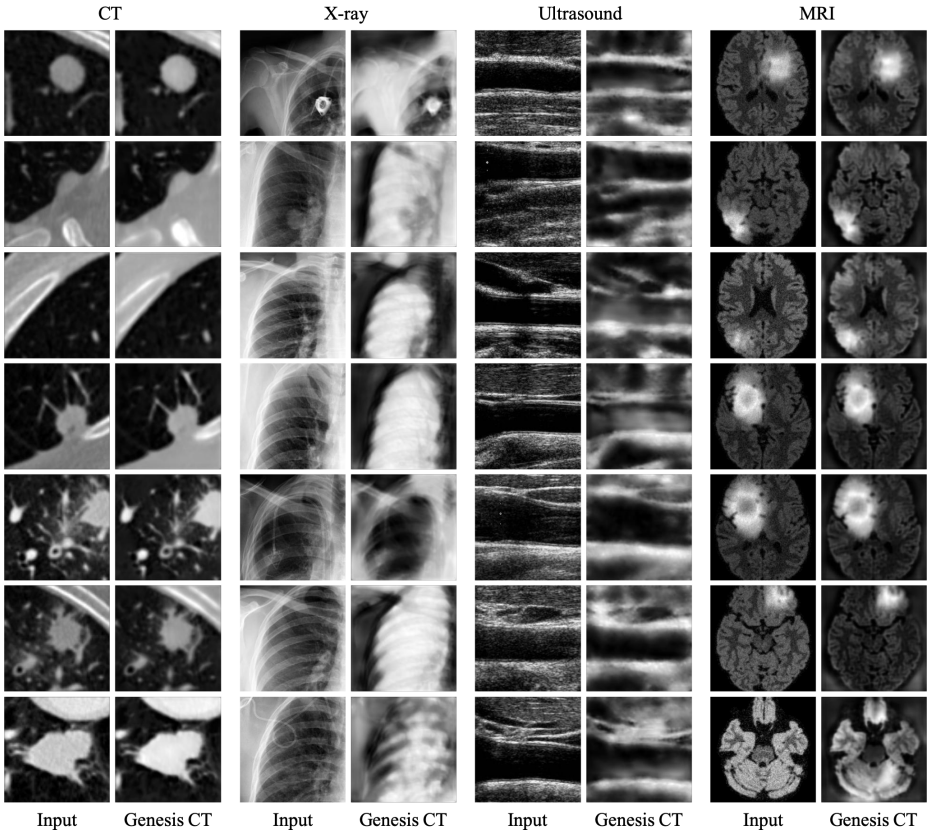


Fig. 11: Qualitative assessment of image restoration quality by Genesis Chest CT across dataset, organ, and modality is visualized. Genesis Chest CT is trained on LIDC-IDRI (CT) [1] via our unified self-supervised training framework. For testing, we use the pre-trained model to directly restore images from LIDC-IDRI (CT), ChestX-ray8 (X-ray) [17], CIMT (Ultrasound) [8,18], and BraTS (MRI) [13]. Though the model is only trained on CT data, it can largely maintain the texture and structures during restoration not only in the same modality (CT), but also in different modalities including X-ray, Ultrasound, and MRI, suggesting that Genesis Chest CT is transferable across datasets, organs, and modalities. Besides, we notice that the restoration quality is also consistent with the results of Genesis Chest CT on target tasks (see Fig. 2). For example, compared to cross-modality performance, Genesis Chest CT yields better performance in CT for both restoration and target tasks (*i.e.,* `NCC` and `NCS`). Moreover, the relative lower restoration quality of ultrasound images may explain the relative lower target performance of Genesis Chest CT on `IUC` (see Fig. 2). Finally, by comparing the performance of Genesis Chest CT in various modalities, we find out that a model pre-trained in the same domain is still preferred whenever possible. Thereby, we will continue developing modality-oriented models including Genesis MRI and Genesis Ultrasound.

G Genesis Chest X-ray

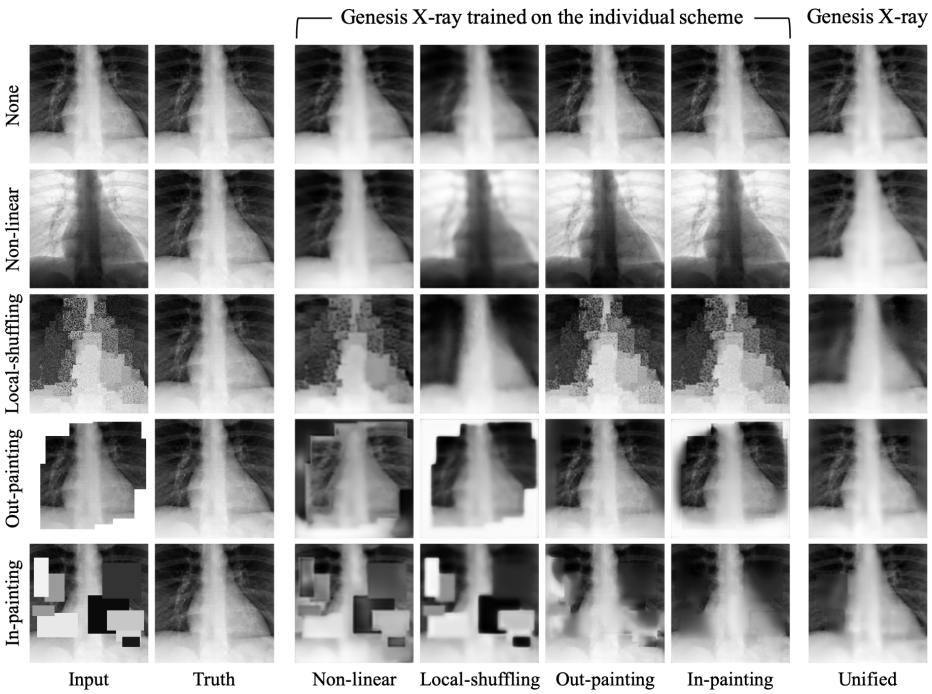


Fig. 12: Qualitative comparisons of Genesis Chest X-ray trained with unified self-supervised framework and four individual training schemes. We train and test all five models on ChestX-ray8 [17] where transformed patches (Column 1) undergo one of the four transformations (Rows 2—5) as well as an identity mapping (Row 1). It is clear from the figure that the models trained with a single transformation fail to handle other transformations. For example, considering the training scheme based on in-painting (Row 5), models trained on individual training schemes fail to in-paint the masked region except for the in-painting-trained model (Column 6). However, the model trained with the unified framework (Column 7) handles all of the transformations and generates patches fairly close to the ground truths. These observations suggest that Models Genesis trained with proposed unified self-supervised learning framework learns general anatomical structures better, yielding high-performance target models through transfer learning.

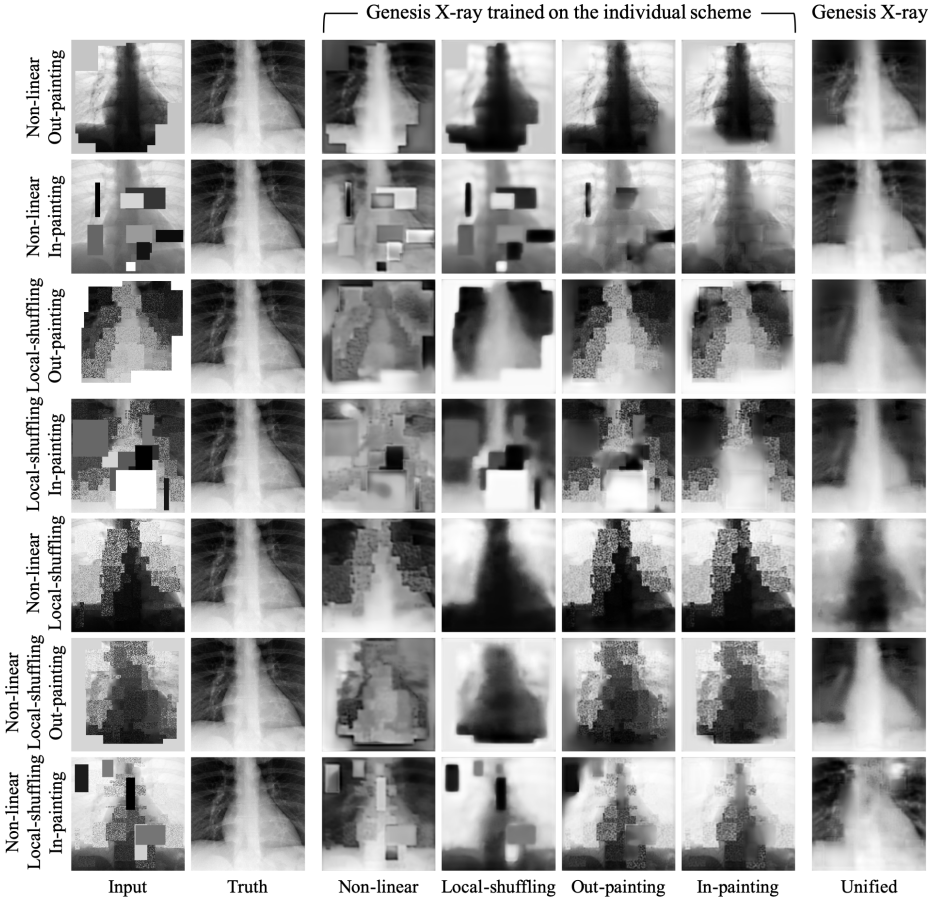


Fig. 13: Continued from Fig. 12. We further test our five models on seven combinations of transformations. The models train with individual training scheme can only handle a single transformation (Columns 3—6) and fail to restore the patches completely, while Models Genesis trained via proposed unified self-supervised learning framework (Column 7) fairly handle seven augmented transformations and restores the patches close to the original patch. Taking a combination of out-painting and non-linear transformation (Row 1) as an example, the model trained on non-linear-based scheme (Column 3) recovers the original intensity values, but fails to out-paint the image; however, the model trained with a unified framework not only recovers the original intensity values but also out-paints the image. This observation demonstrates the superiority of Models Genesis trained with unified self-supervised learning framework.

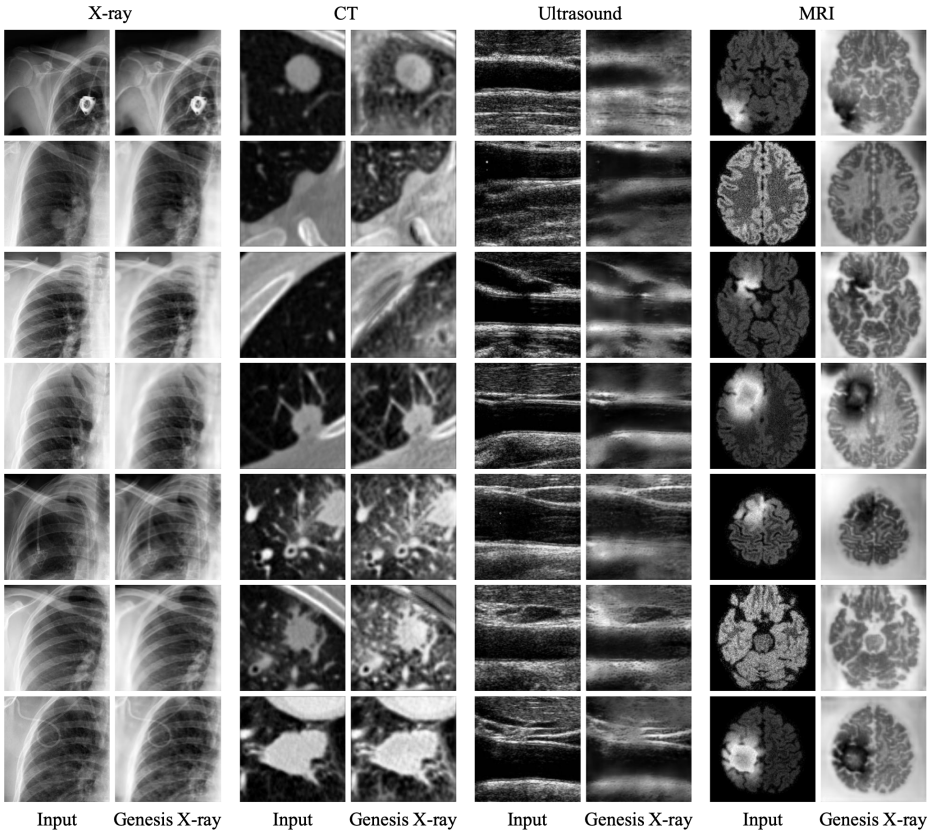
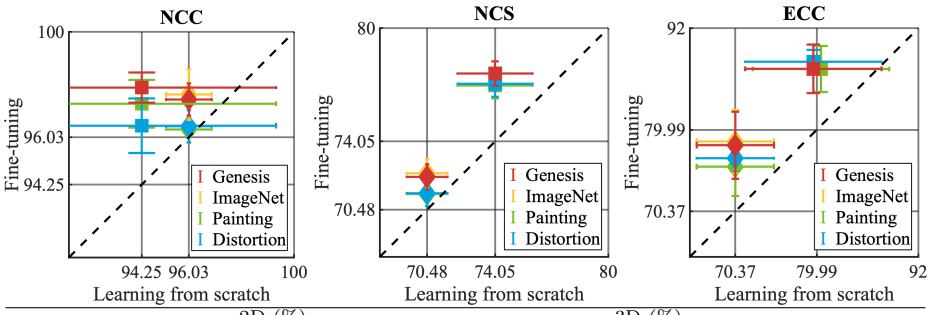


Fig. 14: Qualitative results of image restoration from Genesis Chest X-ray across dataset, organ, and modality are visualized. Genesis Chest X-ray is trained on Chest X-ray8 (X-ray) [17] via our unified training framework, and tested to restore images from Chest X-ray8 (X-ray), LIDC-IDRI (CT) [1], CIMT (Ultrasound) [8,18], and BraTS (MRI) [13]. Similar to Fig. 11, we observe that the performance of restoration and target tasks in various modalities may be positively correlated. For instance, while Genesis Chest X-ray restores ultrasound images reasonably, it injects unintended artifacts in the restored CT slices. As a result, Genesis Chest X-ray achieves better performance on `IUC` task compared with Genesis Chest CT, but it fails on `NCC` task (see Fig. 2). The analysis further confirms our claims provided in Fig. 11.

H Models Genesis vs. Models ImageNet



Task	2D (%)			3D (%)			p-value†
	Scratch	ImageNet	Genesis	Scratch	ImageNet	Genesis	
NCC	96.03±0.86	97.79±0.71	97.45±0.61	94.25±5.07	N/A	**98.20±0.51**	0.0213
NCS	70.48±1.07	72.39±0.77	72.20±0.67	74.05±1.97	N/A	**77.62±0.64**	<1e-8
ECC	71.27±4.64	78.61±3.73	78.58±3.67	79.99±8.06	N/A	**88.04±1.40**	5.50e-4

†These p-values are calculated between our Models Genesis vs. the fine-tuning from ImageNet, which always offers the best performance (highlighted in red) for all three tasks in 2D.

Approach	NCC (%)	NCS (%)	ECC (%)	LCS (%)	BMS (%)
Scratch	94.25±5.07	74.05±1.97	79.99±8.06	74.60±4.57	90.16±0.41
Distortion (ours)	96.46±1.03	77.08±0.68	**88.04±1.40**	79.08±4.26	**90.60±0.20**
Painting (ours)	**98.20±0.51**	77.02±0.58	87.18±2.72	78.62±4.05	90.46±0.21
Unified (ours)	97.90±0.57	**77.62±0.64**	87.20±2.87	**79.52±4.77**	90.59±0.21
p-value††	0.0848	0.0520	0.2102	0.4249	0.4276

††These p-values are calculated between the top-2 models in each column highlighted in red.

Fig. 15: Comparison of Models Genesis and Models ImageNet. In the top three sub-figures, the 3D volume-based solutions and 2D slice-based solutions are denoted with square and diamond markers, respectively. The horizontal and vertical error bars indicate 95% confidence intervals of training from scratch and fine-tuning, respectively. The shorter the vertical bar, the more consistent and stable the model is.

The comparisons of our Models Genesis and Models ImageNet (*i.e.,* models pre-trained on ImageNet) are summarized in three figures and two tables in Fig. 15. Training 3D models simply from scratch does not necessarily outperform the 2D counterparts (see NCC), however, fine-tuning the same 3D models from Genesis Chest CT significantly outperforms ($p < 0.05$) the slice-based 2D models including fine-tuning from Models ImageNet. As seen, Models Genesis enjoys a higher stability on the target tasks. Moreover, comparing our unified framework with individual training schemes demonstrates that the former is more robust across all target tasks, yielding either the best result or comparable performance to the best model ($p < 0.05$). This superiority of our Models Genesis is attributable to consolidating multiple self-supervised training schemes, which enables the model to learn a stronger image representation. Thus, fine-tuning Models Genesis leads to powerful and stable application-specific target models, confirming the importance of Models Genesis in 3D medical imaging.

I The NiftyNet Transfer Learning Capability

Initialization	NCS (Dice %)	LCS (Dice %)	BMS (Dice %)
Models Genesis (ours)	75.86±0.90	91.13±1.51	92.58±0.30
NiftyNet scratch [7]	69.65±2.56	91.09±0.76	90.68±0.24
NiftyNet model zoo [6]	69.24±1.77	90.84±0.63	90.65±0.54
p-value†	0.3433	0.2214	0.4301

†These p-values are calculated between NiftyNet scratch and NiftyNet model zoo.

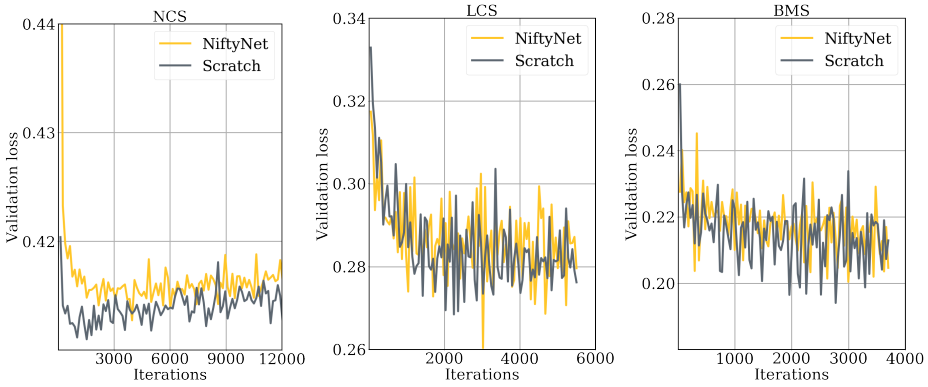


Fig. 16: Fine-tuning the pre-trained NiftyNet vs. training it from scratch. The results reported in the table statistically infer that fine-tuning the pre-trained NiftyNet offers no benefit over training it from scratch. This fact is further supported by the learning curve comparison provided in the figures.

The table in Fig. 16 compares fine-tuning the pre-trained NiftyNet with training from scratch on three target tasks: (1) lung nodule segmentation (NCS) in CT images, (2) liver segmentation (LCS) in CT images, and (3) brain tumor segmentation (BMS) in MRI images using dice-coefficient (mean±s.d.) as the evaluation metric, demonstrating that fine-tuning NiftyNet's 3D supervised pre-trained weights has *no* benefit over random initialization ($p > 0.05$). It is further corroborated by the learning curves on validation dataset provided at the bottom in Fig. 16. However, Models Genesis significantly improve performance over random initialization (see Table 2 in the main paper) and perform consistently better than NiftyNet models on the three same target tasks. Note that the pre-trained NiftyNet model was trained using strong supervision, whereas Models Genesis learns representations using the proposed self-supervised paradigm. In contrast to pre-trained weights of NiftyNet's model zoo, the pre-trained weights from our proposed self-supervised method are found to be more robust across diseases, organs, and imaging modalities, thanks to the ability of our approach to learn representations from a large-scale unannotated dataset.

References

1. Armato III, S.G., *et al.*: 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(2), 915–931 (2011) 19, 21, 24
2. Buzug, T.M.: Computed tomography. In: Springer Handbook of Medical Technology, 311–342 (2011) 13
3. Chen, T., *et al.*: Self-Supervised GANs via Auxiliary Rotation Loss. In: CVPR, 12154-12163 (2019) 12
4. Deng, J., *et al.*: ImageNet: A large-scale hierarchical image database. In: CVPR, 248–255 (2009) 12
5. Forbes, G.B.: Human body composition: growth, aging, nutrition, and activity. Springer Science & Business Media (2012) 13
6. Gibson, E., *et al.*: Automatic multi-organ segmentation on abdominal ct with dense v-networks. TMI, 37(8), 1822–1834 (2018) 26
7. Gibson, E., *et al.*: Niftynet: a deep-learning platform for medical imaging. Computer Methods and Programs in Biomedicine (2018), https://www.sciencedirect.com/science/article/pii/S0169260717311823 26
8. Hurst, R.T., *et al.*: Incidence of subclinical atherosclerosis as a marker of cardiovascular risk in retired professional football players.The American journal of cardiology, 105(8), 1107–1111 (2010) 21, 24
9. Iizuka, S., *et al.*: Globally and locally consistent image completion. ACM Transactions on Graphics (ToG), 36(4), 107 (2017) 16
10. Jing, L., *et al.*: Self-supervised visual feature learning with deep neural networks: A survey. arXiv:1902.06162 (2019) 12
11. Krizhevsky, A., *et al.*: Imagenet classification with deep convolutional neural networks. In: NIPS, 1097–1105 (2012) 12
12. Kolesnikov, A., *et al.*: Revisiting self-supervised visual representation learning. In: CVPR, 1920–1929 (2019) 12
13. Menze, B.H., *et al.*: The multimodal brain tumor image segmentation benchmark (brats). TMI, 34(10), 1993 (2015) 21, 24
14. Mortenson, M.E.: Mathematics for computer graphics applications. Industrial Press Inc. (1999) 13
15. Pathak, D., *et al.*: Context encoders: Feature learning by inpainting. In: CVPR, 2536–2544 (2016) 16
16. Vincent, P., *et al.*: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research 11(Dec), 3371–3408 (2010) 16
17. Wang, X., *et al.*: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR, 2097–2106 (2017) 21, 22, 24
18. Zhou, Z., *et al.*: Integrating active learning and transfer learning for carotid intima-media thickness video interpretation. Journal of digital imaging, 32(2), 290-299 (2019). 21, 24