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)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-32251-9_42) contains supplementary material, which is available to authorized users.

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11767, pp. 384–393, 2019.
https://doi.org/10.1007/978-3-030-32251-9_42

Table 1. Target tasks.

Code[a]	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

[a] 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).

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 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 Sect. 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 Ima-

[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.

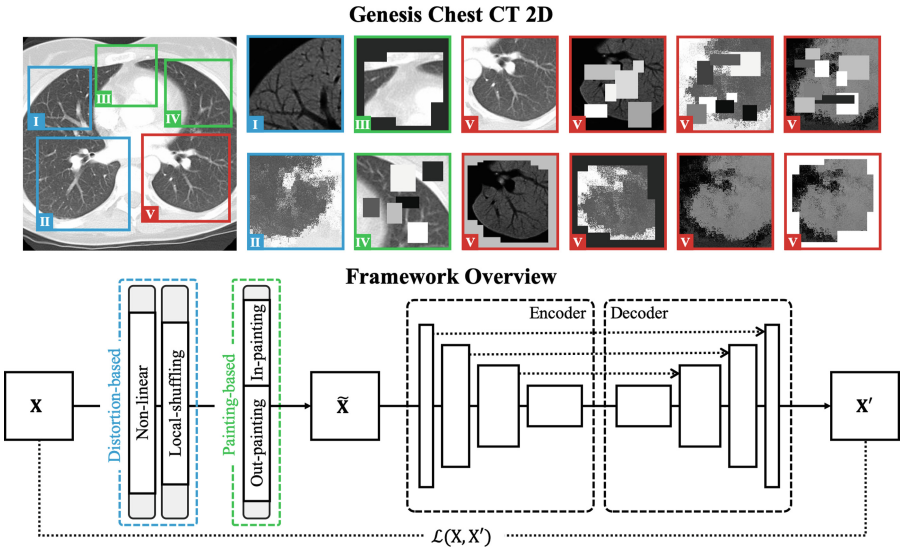

Fig. 1. Our unified self-supervised learning framework consolidates four novel transformations: (I) non-linear, (II) local-shuffling, (III) out-painting, and (IV) in-painting into a single image restoration task. Specifically, each arbitrarily-size patch X cropped at random location from an unlabeled 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.

geNet 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). 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] Sect. 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 (See footnote 4) Sect. 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

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

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 (See footnote 4) Sects. 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 modest benefit to 3D medical imaging applications as well as that from the pre-trained models of NiftyNet[5], which is ineffective (see Sect. 3 and Appendix (See footnote 4) Sect. 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

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

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 (See footnote 1) and 77,074 X-rays in ChestX-ray8(See footnote 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 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

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

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[a]	AUC					94.25 ± 5.07	$\mathbf{98.20 \pm 0.51}$	0.0180
NCS[b]	IoU					74.05 ± 1.97	$\mathbf{77.62 \pm 0.64}$	1.04e–4
ECC[c]	AUC	✗		✗		79.99 ± 8.06	$\mathbf{88.04 \pm 1.40}$	0.0058
LCS[d]	IoU	✗	✗	✗		74.60 ± 4.57	$\mathbf{79.52 \pm 4.77}$	0.0361
BMS[e]	IoU	✗	✗	✗	✗	90.16 ± 0.41	$\mathbf{90.60 \pm 0.20}$	0.0041

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

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 bold and italic. 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*	$\mathbf{88.04 \pm 1.40}$	*79.08 ± 4.26*	$\mathbf{90.60 \pm 0.20}$
Painting (ours)	$\mathbf{98.20 \pm 0.51}$	77.02 ± 0.58	87.18±2.72	78.62 ± 4.05	90.46±0.21
Unified (ours)	*97.90 ± 0.57*	$\mathbf{77.62 \pm 0.64}$	*87.20 ± 2.87*	$\mathbf{79.52 \pm 4.77}$	*90.59 ± 0.21*
p-value	0.0848	0.0520	0.2102	0.4249	0.4276

8 points. Genesis Chest CT continues to yield significant IoU gain for LCS and BMS even though 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

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[a]
	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	$5.50e{-}4$

[a]These p-values are calculated between our Models Genesis vs. the fine-tuning from ImageNet, which always offers the best performance for all three tasks in 2D.

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 (See footnote 4) Sect. 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.

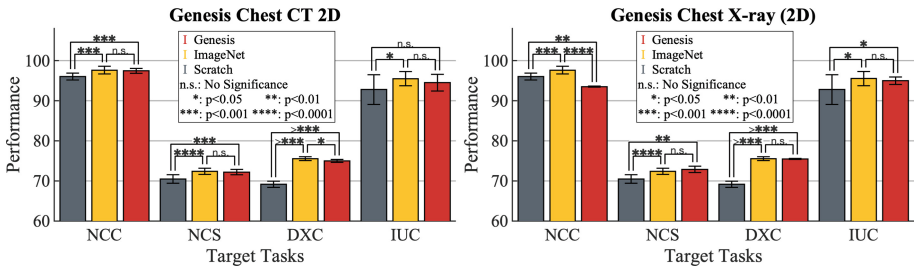

Fig. 2. Comparison of 2D solutions on four 2D target tasks. To investigate the same- and cross-domain transferability of Models Genesis, we have trained Genesis Chest CT 2D using 2D axial slices from LUNA dataset (left panel), and Genesis Chest X-ray (2D) trained using radiographs from ChestX-ray8 dataset (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.

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 (See footnote 4) Sects. 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 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 (See footnote 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 (See footnote 4) Sect. 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)
2. Doersch, C., et al.: Multi-task self-supervised visual learning. In: ICCV **2051–2060**, (2017)
3. Jamaludin, A., Kadir, T., Zisserman, A.: Self-supervised learning for spinal MRIs. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 294–302. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_34
4. Jing, L., et al.: Self-supervised visual feature learning with deep neural networks: A survey. arXiv:1902.06162 (2019)
5. Kang, G., et al.: Patchshuffle regularization. arXiv:1707.07103 (2017)
6. Pathak, D., et al.: Context encoders: Feature learning by inpainting. In: CVPR, 2536–2544 (2016)
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)
8. Spitzer, H., et al.: Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In: MICCAI, 663–671 (2018)
9. Tajbakhsh, N., et al.: Convolutional neural networks for medical image analysis: Full training or fine tuning? TMI **35**(5), 1299–1312 (2016)
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)