Parts2Whole: Self-supervised Contrastive Learning via Reconstruction

Ruibin Feng[1(✉)], Zongwei Zhou[1], Michael B. Gotway[2], and Jianming Liang[1(✉)]

[1] Arizona State University, Tempe, AZ 85281, USA
{rfeng12,zongweiz,jianming.liang}@asu.edu
[2] Mayo Clinic, Scottsdale, AZ 85259, USA
Gotway.Michael@mayo.edu

Abstract. Contrastive representation learning is the state of the art in computer vision, but requires huge mini-batch sizes, special network design, or memory banks, making it unappealing for 3D medical imaging, while in 3D medical imaging, reconstruction-based self-supervised learning reaches a new height in performance, but lacks mechanisms to learn contrastive representation; therefore, this paper proposes a new framework for self-supervised contrastive learning via reconstruction, called Parts2Whole, because it exploits the *universal* and *intrinsic* part-whole relationship to learn contrastive representation without using contrastive loss: Reconstructing an image (whole) from its own parts compels the model to learn similar latent features for all its own partsin the latent space, while reconstructing different images (wholes) from their respective parts forces the model to simultaneously push those parts belonging to different wholes farther apart from each other in the latent space; thereby the trained model is capable of distinguishing images. We have evaluated our Parts2Whole on five distinct imaging tasks covering both classification and segmentation, and compared it with four competing publicly available 3D pretrained models, showing that Parts2Whole significantly outperforms in two out of five tasks while achieves competitive performance on the rest three. This superior performance is attributable to the contrastive representations learned with Parts2Whole. Codes and pretrained models are available at github.com/JLiangLab/Parts2Whole.

Keywords: 3D Self-supervised Learning · Contrastive representation learning · Transfer learning

1 Introduction and Related Work

Contrastive representation learning has made a leap in computer vision. For example, MoCo [13] introduces the momentum mechanism, and SimCLR [10] proposes a simple framework for contrastive learning; both methods achieve state-of-the-art results and even outperform supervised ImageNet pretraining. However, contrastive learning requires huge mini-batch sizes [10,14], special network design [3], or memory banks [13,14,19] to store feature representations of

S. Albarqouni et al. (Eds.): DART 2020/DCL 2020, LNCS 12444, pp. 85–95, 2020.
https://doi.org/10.1007/978-3-030-60548-3_9

all images in the dataset, making it unattractive for 3D medical imaging applications. Taking the mini-batch size as an example, SimCLR [10] recommends 8192, which is impractical for 3D image data due to the current GPU memory limitation. On the other hand, reconstruction-based self-supervised learning has proven to be effective and efficient for 3D medical image analysis. Models Genesis [20] establish autodidactic models by restoring images that underwent four transformations. Later, Tao *et al.* [18] permute volumetric data via 3D voxel rotation and then restore the original data to learn robust features. Therefore, in this paper, we seek to answer the following critical question: *Can we learn contrastive representations via reconstruction for 3D medical imaging to effectively address the aforementioned barriers associated with contrastive learning?*

To answer this question, we exploit a *universal* and *intrinsic* property, the part-whole relationship, where an entire image is regarded as the whole and any of its patches are considered as its parts. This property has been explored in SimCLR [10] via contrastive prediction between the global view (whole) and local view (part). Later, SwAV [7] observed that mapping local views to global views can significantly increase the representation quality. However, instead of directly comparing features or their cluster assignments, we reconstruct a whole from its parts with a pair of encoder and decoder. By doing so, the deep model is compelled to learn contrastive representations embedded with part-whole semantics: (1) the representations of parts belonging to the same whole are close, and (2) the representations of parts belonging to different wholes are far away. We refer to our self-supervised learning framework as Parts2Whole.

Notably, Parts2Whole integrates advantages of several existing self-supervised learning approaches, but overcomes their limitations: Parts2Whole (1) discriminates individual images as Exemplar [11] aims for,[1] but overcomes the scalability issue stemmed from classification because of our use of reconstruction; (2) learns contrastive representations like contrastive learning methods [10,13], but eliminate the need for huge mini-batch size or memory bank by avoiding direct feature comparison; (3) exploits recurrent anatomical structures like Models Genesis [20], but enriches feature representations with part-whole semantics. Furthermore, Models Genesis [20], particularly in-painting/out-painting, only interpolates/extrapolates known masked regions since there are pixel-coordinate mappings between inputs and ground truths. However, in Parts2Whole, the restored area is unknown and the pixel-coordinate mapping does not exist because inputs are randomly cropped and resized from ground truths. In other words, Parts2Whole needs to learn the scale of wholes as well as recover the missing contents, which is much harder and therefore yields more powerful source models.

Our pretrained model is extensively evaluated on five distinct medical target tasks and compare with four competing publicly available 3D models pretrained in either fully supervised or self-supervised fashion in Sect. 3.2. The statistical

[1] If we consider each whole image itself as a "label", the training process of Parts2Whole is equivalent to predicting the correct "label" given a part of one image as input, or discriminating each image from its parts.

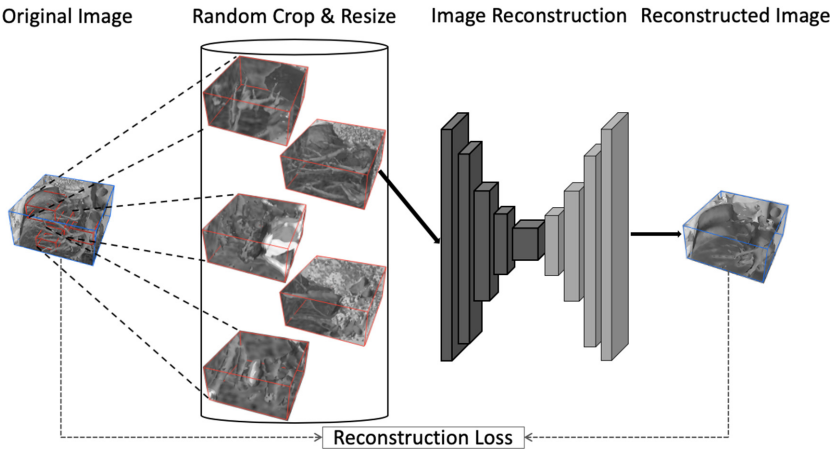

Fig. 1. We propose a new self-supervised learning framework, called Parts2Whole, because it exploits a *universal* and *intrinsic* property, the part-whole relationship, where an entire image is regarded as the whole and its cropped patches are considered as its parts. Parts2Whole aims to learn contrastive representations that embed the part-whole semantics by reconstructing a whole (blue framed) from its resized randomly cropped parts (red framed). To avoid trivial solutions, we crop each whole with random scales and aspect ratios to erase low-level cues across different parts while maintaining informative structures and textures. Additionally, we do not use the skip connections to avoid low-level details passing from the encoder to decoder, yielding generic pretrained models with strong transferability. The model is trained in an end-to-end fashion and the reconstruction loss is measured with Euclidean distance. (Color figure online)

analysis in Table 1 shows Parts2Whole significantly outperforms in two out of five tasks while achieves competitive performance on the rest three. We also empirically validate that Parts2Whole can learn contrastive representations in an image reconstruction framework in Fig. 2 and Sect. 3.3. Finally, our Parts2Whole design is justified by ablating its main components in Table 2.

2 Proposed Method

Our goal is to learn contrastive representations embedded with part-whole semantics by reconstructing the whole image from its parts, with 3D unlabeled images. The proposed self-supervised learning framework is illustrated in Fig. 1.

Problem Formulation: Denote a set of 3D unlabeled images as $\{x_i \in X : i \in [1, N]\}$ where N is the number of whole images. Each image x_i is randomly cropped and resized to generate various parts, referred to as $\{p_i^j \in P_i : i \in [1, N], j \in [1, M)\}$. The task is to predict the whole image x_i from its local patch p_i^j by training a pair of encoder ($\mathcal{F}_E$) and decoder ($\mathcal{F}_D$) to minimize the loss function, denoted by $\mathcal{L} = \sum_i \sum_j l(\mathcal{F}_D(\mathcal{F}_E(p_i^j)), x_i)$, where $l(\cdot)$ is a metric

measuring the difference between the model outputs and ground truths. We use Euclidean distance as $l(\cdot)$ in this work.

Since the output images are generated via a *shared* decoder ($\mathcal{F}_D$), the encoder ($\mathcal{F}_E$) is forced to learn contrastive representations that embed the part-whole semantics. To be specific, after training, $\mathcal{F}_E(p_i^j)$ and $\mathcal{F}_E(p_{i'}^{j'})$ are forced to be as close as possible if $i = i'$ since the two representations are mapped to the same groud truth (x_i) via the shared decoder ($\mathcal{F}_D$), while far away from each other otherwise since they are mapped to different ground truths. To avoid ambiguous cases, we assume that *no part is also a whole*.

Removing Skip Connection: The skip connection (or the shortcut connection) was proposed in [6] and adopt to connect the encoder and decoder in the U-Net architecture [15]. The goal is to let the decoder access the low-level features produced by the encoder layers such that the boundaries in segmentation maps produced by the decoder could be accurate. However, we argue that if the network can solve the proxy task using lower-level cues, it does not need to learn semantically meaningful representations. Therefore, in proxy task training, we adopt the 3D U-Net architecture[2] and remove the skip connections to force the bottleneck representations encoding high-level information, which is different from [18,20] in the perspective of network architectures. Table 2 demonstrates the effects of skip connections in proxy task training. Notice that although we cannot provide a pretrained decoder as [20] does, our model offers very competitive performance on three segmentation tasks with a randomly initialized decoder, suggesting our pretrained encoder learns strong, generic features.

Extracting Local Yet Informative Parts: The part size is a critical component in our proxy task design. For example, when the crop scale is too large, the task is downgraded to training an autoencoder without learning semantics. On the other hand, the task can be unsolvable if the parts are too small and do not contain enough information. To avoid such degenerate solutions, we restrict the cropped patches covering less than 1/4 area of the whole image. By doing so, the low-level cues across different parts are largely erased. Additionally, we set each part covering more than 1/16 area of the original image to have discriminative structures and textures. We analyze the effects of crop scales in Table 2.

3 Experiments

3.1 Experiment Settings

Proxy Task Training: We pretrain our model on LUNA-2016 [16] dataset without using any label shipped with it. To avoid test data leakage, we use 623 CT scans instead of all 888 scans. We first cropped the original CT scans to small, *non-overlapped* 28,144 sub-scans with dimensions equal to 128×128×64.

[2] 3D U-Net: github.com/ellisdg/3DUnetCNN.

Table 1. Our pretrained model achieves significantly better or at least comparable performance on five distinct medical target tasks over four publicly available 3D models pretrained in both supervised and self-supervised fashion. Each experiment is conducted for 10 trials and summarized with the mean and standard deviation (mean±s.t.d.). The paired t-test results between our method and the previous top-1 solution are tabulated in terms of the p-value. The best approaches are **bolded** while the others are highlighted in blue if they achieve equivalent performance compared with the best one (*i.e.,* $p > 0.05$).

Approach	NCC AUC(%)	NCS IoU(%)	LCS†† IoU(%)	ECC AUC(%)	BMS††† IoU(%)
Scratch	94.25±5.07	74.05±1.97	77.82±3.87	79.99±8.06	63.91±1.41
I3D [8]	98.26±0.27	71.58±0.55	70.65±4.26	80.55±1.11	67.83±0.75
NiftyNet [12]	94.14±4.57	52.98±2.05	83.23±1.05	77.33±8.05	60.78±1.60
MedicalNet [9]	95.80±0.49	75.68±0.32	85.52±0.58	86.43±1.44	66.09±1.35
Models Genesis [20]	97.90±0.57	**77.62±0.64**	84.17±1.93	**87.20±2.87**	68.08±1.15
Parts2Whole (ours)	**98.67±0.23**	77.35±0.61	**86.70±0.62**	86.14±2.97	**68.33±0.41**
p-value†	0.0011	0.1709	0.0002	0.2126	0.2654

† p-values are calculated between Parts2Whole and the previous top-1 solution.
†† The IoU score is calculated using binarized masks with a threshold equal to 0.5 to better presented the segmentation quality, while [20] uses the original masks without thresholding.
††† Notice the results are different from those reported in [20] since we use *real* data while Models Genesis were evaluated with *synthetic* data.

We treat each generated sub-scan as a *whole* and crop parts from it on the fly. The cropped parts contain [1/16, 1/4] volume of the whole image.

Target Task Training: To extensively evaluate our pretrained 3D model, we follow the practice employed in [20] and investigate five distinct medical applications, including lung nodule false positive reduction (NCC) [16], lung nodule segmentation (NCS) [2], liver segmentation (LCS) [5], pulmonary embolism false positive reduction (ECC) [17], and brain tumor segmentation (BMS) [4].

3.2 Parts2Whole Yields Competitive 3D Pretrained Models

To extensively evaluate our method, we compare Parts2Whole with four *publicly available 3D models* pretrained in both supervised and self-supervised fashion. To be specific, we test two models supervisely pretrained on 3D medical segmentation tasks: NiftyNet [12] with Dense V-Networks and MedicalNet [9] with 3D-ResNet-101 as the backbone. The former is pretrained with a multi-organ CT segmentation task, and the latter is pretrained with an aggregate dataset (*i.e.,* 3DSeg-8) from eight public medical datasets. We also evaluate I3D [8], which is pretrained with natural videos but has been successfully applied for lung cancer classification [1]. For self-supervised learning, we choose Models Genesis [20], the current state of the art in 3D medical imaging, as our baseline.

The experimental results are summarized in Table 1. First of all, we observe that I3D works well on NCC but performs inferiorly on the other 4 tasks. This suboptimal performance may attribute to the marked difference between natural and medical domains. On the other hand, NiftyNet and MedicalNet, which are fully supervised with medical data, also show relatively poor transferability. We hypothesize that the main reason is the limited amount of annotation for supervising. A piece of evidence is that MedicalNet considerably outperforms NiftyNet by aggregating eight datasets for pretraining. These observations highlight the significance of self-supervised learning in the 3D medical domain, which can close the domain gap and utilize the vast amount of unannotated data.

In contrast with fully supervised pretraining, both self-supervised learning methods (Models Genesis and Parts2Whole) achieve promising results on all five target tasks across organs, diseases, datasets, and modalities. Specifically, for NCC and LCS, Parts2Whole not only has higher AUC/IoU scores and lower standard deviations but also significantly outperforms Models Genesis based on the t-test ($p < 0.05$). On the other hand, Models Genesis achieves better performance by a small margin on NCS and ECC tasks. On the BMS task, which has considerable distance from the proxy dataset (*i.e.,* different disease, organ, and modality), Parts2Whole is still competitive compared to other baselines. Last but not least, since Models Genesis provides both pretrained encoder and decoder, one can expect it to have certain advantages on segmentation tasks (*i.e.,* NCS, LCS, and BMS). Nonetheless, Parts2Whole yields promising results on all segmentation tasks with the same architecture (*i.e.,* 3D U-Net) and a randomly initialized decoder, suggesting the encoder pretrained with Parts2Whole learns features with strong transferability. Next, we will experimentally investigate the properties of feature representations learned in Parts2Whole.

3.3 Parts2Whole Learns Contrastive Representations

To understand the feature representations learned with Parts2Whole, we first visualize the t-SNE embeddings of random, Models Genesis, and Parts2Whole features in Fig. 2(a). Specifically, we randomly select 10 whole images and generate 200 parts for each image with a crop scale [1/16, 1]. Each circle represents one part while each diamond represents a whole image (*i.e.,* crop scale is equal to 1). Different colors and circle sizes denote different images and crop scales.

First of all, unlike the entangled features from random initialization, features pretrained with Models Genesis and Parts2Whole are more separable. However, Models Genesis features are not distinguishable especially when the inputs are cropped with small scales (seeing the red-framed part). On the contrary, Parts2Whole features from the same image are well grouped, while those from different images are highly separable regardless of different crop scales. More importantly, although the network is never trained with large patches or the whole image (*i.e.,* crop scale is equal to [1/4, 1]), it correctly aligns all features from the same image together. This magnificent generalization ability suggests that Parts2Whole learns representations that embed the part-whole semantics.

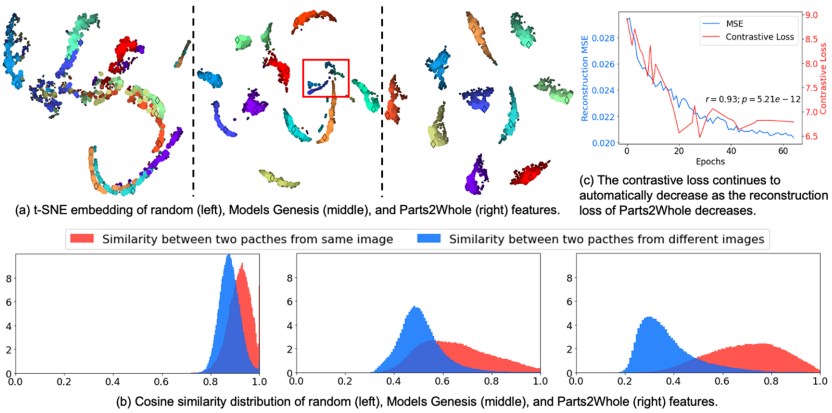

Fig. 2. To understand the learned representations, we first visualize the t-SNE embeddings of random, Models Genesis, and Parts2Whole features in (a). We use circles to represent parts while diamonds represent whole images. The colors and circle sizes denote the different wholes and crop scales. Compared with random and Models Genesis features, the Parts2Whole features from the same images are well grouped, while features from different images are highly separable, despite the different crop scales. Furthermore, we leverage the entire validation set and measure the cosine similarity between features of two parts belonging to the same or different images in (b). Notice that the similarity distributions of Parts2Whole features are more separable than those of random and Models Genesis features, indicating that Parts2Whole learns better representations. Last but not least, as shown in (c), the contrastive loss continues to automatically decrease validating Parts2Whole can learn contrastive representations.

To further analyze the feature representations, we leverage all the validation images. For each image x_i, we generate 100 pairs of parts from it (referred to as positive pairs). Additionally, we generate 100 negative pairs, while each one contains one part from x_i and one part from another random picked image. We calculate the cosine similarity between each positive pair and negative pair. The similarity distributions of random, Models Genesis, and Parts2Whole features are shown in Fig. 2(b). Note that the similarity distributions of Parts2Whole features are more separable than those of Models Genesis and random features. It further indicates that Parts2Whole learns contrastive representations.

We also investigate the change of the contrastive loss along the training process in Fig. 2(c). We test every whole image 100 times, while in each test, we randomly generate 1 positive pair and 5000 negative pairs for the contrastive loss[3] calculation. As illustrated, the contrastive loss continues to decrease as the reconstruction loss decreases. Additionally, we perform the Pearson product-moment correlation analysis between the reconstruction loss and the contrastive loss.

[3] Denote the l_2-normalized features of a positive pair and negative pair as $\{\mathcal{F}_E(p_i), \mathcal{F}_E(p_i')\}$ and $\{\mathcal{F}_E(p_i), \mathcal{F}_E(p_j)\}$, respectively. The contrastive loss is calculated as $-\log \frac{\exp(\mathcal{F}_E(p_i)\cdot\mathcal{F}_E(p_i')/\tau)}{\exp(\mathcal{F}_E(p_i)\cdot\mathcal{F}_E(p_i')/\tau+\sum_{j=1}^{5000}\exp(\mathcal{F}_E(p_i)\cdot\mathcal{F}_E(p_j)/\tau)}$ where $\tau = 0.7$.

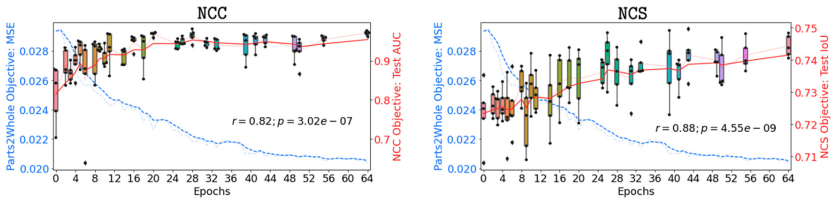

Fig. 3. The overall performance of target tasks continues to improve as the validation loss in the proxy task decreases. We validate the consistency of proxy and `NCC`/`NCS` target objective by evaluating 26 checkpoints saved in the proxy training process. It is clear that as the proxy loss decreases, the average AUC/IoU score increases while the standard deviation decreases, suggesting that the pretrained model becomes more generic and robust. Additionally, the Pearson product-moment correlation analysis indicates a strong *positive* co-relationship between proxy and target objectives (Pearson's r-value > 0.5).

The high Pearson's r-values (0.93) suggest a strong positive correlation, validating that Parts2Whole can minimize the contrastive loss and learn contrastive representations with an image reconstruction framework. *Notice that we achieve the goal of contrastive learning with small mini-batch sizes (16 instead of 8192 suggested in* [10]), a general 3D U-Net architecture, and without using memory banks — effectively addressing the barriers associated with previous contrastive learning methods. However, it is still not clear whether good contrastive features embedded with part-whole semantics can yield strong transferability, since the proxy task is agnostic about the target tasks. To answer this question, we systematically investigate the relationship between the reconstruction loss in the proxy task and the test performance in target tasks in the next section.

3.4 Parts2Whole's Objective Is Positively Correlated with Target Objectives

Wu *et al.* [19] suggested that a good proxy task is able to improve the target task performance consistently as the proxy objective is optimized. Following this practice, we validate the consistency of proxy and task objectives by evaluating 26 checkpoints saved in the proxy training process. Specifically, we fine-tune every checkpoint 5 times on `NCC` and `NCS` target tasks. To reduce the computational cost, we only use partial training data (45% and 10% for `NCC` and `NCS`, respectively. We plot the proxy reconstruction loss and target scores (AUC/IoU) as a function of proxy task training epochs in Fig. 3. We can observe that, as the reconstruction ability in the proxy task improves (*i.e.,* the validation MSE decreases), the transferability of the pretrained model also improves (*i.e.,* the average target score (AUC/IoU) increases while the standard deviation decreases). We further investigate this relationship by performing Pearson product-moment correlation analysis between the proxy objective (*i.e.,* reconstruction quality, measured by (1-MSE)) and target objective (measured by

Table 2. Target task performance on source models pretrained with different proxy task settings. First, removing skip connections (comparing Column 2 to 3) can significantly improve the performance, suggesting skip connections provide shortcuts to solve the proxy task. We also observe that by reducing the cropping scale (from Column 3 to 6), the overall performer continuously increases, plateaus at $[1/16, 1/4]$, and appears to saturate when the scale is less than 1/8. These observations indicate the importance of crop scales in our proxy task design.

Setting	$[\frac{1}{16}, 1]$ w/ s.c.†	$[\frac{1}{16}, 1]$	$[\frac{1}{16}, \frac{1}{2}]$	$[\frac{1}{16}, \frac{1}{4}]$	$[\frac{1}{16}, \frac{1}{8}]$	$[\frac{1}{32}, \frac{1}{16}]$
NCC	88.48±8.24	93.78±2.12	91.48±0.45	**94.84±1.58**	93.52±1.32	91.69±4.12
NCS	70.64±0.21††	72.72±0.42	73.29±0.58	**74.23±0.87**	73.43±0.32	73.66±0.36

† The source model is trained with skip connections (s.c.) between the encoder and decoder.
†† We fine-tune both pretrained encoder and decoder on the target task.

AUC/IoU scores). The high Pearson's r-values (0.82 and 0.88 in NCC and NCS, respectively) suggest a strong *positive* co-relationship between proxy and target objectives. This analysis indicates that our superior target performance is attributable to the decreasing of reconstruction loss and the learned contrastive features.

3.5 Ablation Study

A Good Proxy Task Needs to be Hard But Feasible. Our Parts2Whole design contains two main components: removing skip connections and selecting proper crop scales. We ablate the impacts of the two components to justify our proxy task design. We evaluate source models pretrained with different proxy task settings on NCC and NCS target tasks with 45% and 10% training data, respectively. The experimental results are tabulated in Table 2.

First, we study the effects of skip connections in Column 2 to 3 of Table 2. By removing skip connections while keeping the same cropping scale, the target performance improves significantly by 5.30 and 2.08 points in NCC and NCS, respectively. It suggests that skip connections may pass lower-level details from the encoder to decoder, and ergo provide some shortcuts to solve the proxy task.

Second, with the same network architecture (*i.e.,* no skip connections), we study the effects of different part sizes in Column 3–7 of Table 2. When the upper bound of part sizes is gradually reduced, the overall performance continuously increases, plateaus at 1/4, and appears to saturate at 1/8. On the other hand, when the parts are too small (*i.e.,* less than 1/16), the target performance drops by 3.15 and 0.57 points in NCC and NCS, respectively. These observations indicate the importance of proper part sizes in our proxy task design—the parts should be small enough to avoid trivial solutions while large enough to contain enough information to recover the whole images. In other words, we would like to point out that, naturally, *a good proxy task should be hard enough but still feasible.*

4 Conclusion and Future Work

We present a new self-supervised framework, Parts2Whole, by exploiting the *universal* and *intrinsic* part-whole relationship. Our Parts2Whole can learn contrastive representations in an image reconstruction framework. The experimental results show our pretrained model achieves competitive performance over four publicly available pretrained 3D models on five distinct medical target tasks. However, since we only use the part-whole relationship, incorporating other domain knowledge or transformations may boost the results further, as suggested in [10,20]. One promising direction is to include color/intensity transformations since the similar intensity distribution across parts from one image may provide shortcuts to solve the proxy task [10]. On the other hand, Parts2Whole can minimize contrastive loss without explicitly training with it. It points out an intriguing future work—integrating Parts2Whole and contrastive learning into a unified framework to make the leap in the 3D medical imaging domain.

Acknowledgments. This research has been supported partially by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant, and partially by the 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 would like to thank Jiaxuan Pang, Md Mahfuzur Rahman Siddiquee, and Zuwei Guo for evaluating I3D, NiftyNet, and MedicalNet, respectively. The content of this paper is covered by patents pending.

References

1. Ardila, D., et al.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med. **25**(6), 954–961 (2019)
2. Armato III, S.G., McLennan, G., Bidaut, L., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
3. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. In: Advances in Neural Information Processing Systems, pp. 15509–15519 (2019)
4. Bakas, S., Reyes, M., Jakab, A., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
5. Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., et al.: The liver tumor segmentation benchmark (LiTS). arXiv preprint arXiv:1901.04056 (2019)
6. Bishop, C.M., et al.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)
7. Caron, M., Misra, I., Mairal, J., et al.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020)

8. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
9. Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
11. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 766–774 (2014)
12. Gibson, E., Li, W., Sudre, C., et al.: NiftyNet: a deep-learning platform for medical imaging. Comput. Methods Programs Biomed. **158**, 113–122 (2018)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
14. Misra, I., Maaten, L.V.D.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6707–6717 (2020)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Setio, A.A.A., Traverso, A., De Bel, T., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med. Image Anal. **42**, 1–13 (2017)
17. Tajbakhsh, N., Gotway, M.B., Liang, J.: Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9350, pp. 62–69. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24571-3_8
18. Tao, X., Li, Y., Zhou, W., Ma, K., Zheng, Y.: Revisiting Rubik's cube: self-supervised learning with volume-wise transformation for 3D medical image segmentation. arXiv preprint arXiv:2007.08826 (2020)
19. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
20. Zhou, Z., et al.: Models genesis: generic autodidactic models for 3D medical image analysis. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 384–393. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_42