Supplementary Materials for Anatomical Invariance Modeling and Semantic Alignment for Self-supervised Learning in 3D Medical Image Analysis

Yankai Jiang[1,3*], Mingze Sun[1,4*], Heng Guo[1,2], Xiaoyu Bai[1], Ke Yan[1,2], Le Lu[1] and Minfeng Xu[1,2✉]

[1]DAMO Academy, Alibaba Group
[2]Hupan Lab
[3]College of Computer Science and Technology, Zhejiang University
[4]Tsinghua Shenzhen International Graduate School, Tsinghua-Berkeley Shenzhen Institute, China
✉eric.xmf@alibaba-inc.com

A. Implementation Details

A.1. Preprocessing pipeline for pre-training dataset

The FLARE 2022 dataset is collected from more than 20 medical groups under the license permission, including MSD [16], KiTS [9], AbdomenCT-1K [13], and TCIA [5]. It provides a training set including 2000 unlabelled CT scans with liver, kidney, spleen, or pancreas diseases. We split 10% of the unlabelled CT scans for validation in the pre-training stage, and thus the number of training and validation volumes are 1800 and 200, respectively. **Alice** is pre-trained using only unlabelled images (any annotations were not utilized). First, we clip the CT image intensities from -125 to 255, and then normalize them to the range of 0 to 1. We adopt SAM [20] to locate aligned body parts. The results of landmarks on query and key volumes aligned by SAM are shown in Fig. 5 and Fig. 6. We use a default input volume crop size of $192 \times 192 \times 64$ to generate respective views of consistent anatomies according to the aligned landmarks on each query and key volume pair. In this way, **Alice** is pre-trained via a diverse set of human body compositions, and learns a general-purpose representation from different medical groups' data that can be leveraged for a wide range of downstream tasks.

A.2. End-to-end fine-tuning settings for downstream datasets

We apply our pre-trained online encoder weights to various ViT-based segmentation networks designed for medical tasks of UNETR, nnFormer, and Swin UNETR, by following most of their settings. The detail settings are shown in Tab. 8.

*Equal contribution. This work was done when Yankai Jiang and Mingze Sun were interns at DAMO Academy, Alibaba Group.

B. Results on Downstream Tasks

In this section, we show more results on 3D classification downstream task.

B.1. Dataset

We conduct experiments on a public benchmark MosMedData: Chest CT Scans with COVID-19 Related Findings [14]. This dataset consists of lung CT scans with COVID-19 related findings, as well as without such findings. We use the associated radiological findings of the CT scans as labels and formulate this task as a 2 classes classification to predict presence of viral pneumonia. This dataset contains a total of 1110 CTs. We randomly split 70% of the dataset for training, 10% for validation and the rest 20% for testing. We adopt the ten-fold cross-validation method.

B.2. Preprocessing pipeline

We first rotate the CT volumes by 90 degrees to fix the orientation. We adopt a threshold between -1000 and 400 to clip the CT intensities, and then normalize the Hounsfield units (HU) values to be between 0 and 1. All volumes are resized to $128 \times 128 \times 64$. We use the online data augmentation, including random rotation, scaling, flipping, adding white Gaussian noise, Gaussian blurring, adjusting rightness and contrast, simulation of low resolution.

B.3. Setup

To perform classification, we extract the pretrained online encoder and appended a FC layer with the output chan-

Config	FLARE 2022			BTCV
	UNETR	Swin UNETR	nnFormer	nnFormer
optimizer	AdamW	AdamW	SGD	SGD
base learning rate	$1e^{-4}$	$4e^{-4}$	0.01	0.01
weight decay	$1e^{-5}$	$1e^{-5}$	$3e^{-5}$	$3e^{-5}$
optimizer momentum	0.9	0.9	0.99	0.99
batch size	8	8	8	8
learning rate schedule	cosine decay	cosine decay	"poly" decay	"poly" decay
warmup epochs	50	50	40	40
training epochs	1000	1000	1000	1000
augmentation	random flip, rotation, intensities shifting	random flip, rotation, intensities shifting	scaling, gaussian blur, mirroring	scaling, gaussian blur, mirroring
Spacing	$0.76 \times 0.76 \times 1.5$	$0.76 \times 0.76 \times 1.5$	$0.76 \times 0.76 \times 1.5$	$0.76 \times 0.76 \times 2$
Crop size	$96 \times 96 \times 96$	$96 \times 96 \times 96$	$128 \times 128 \times 96$	$128 \times 128 \times 96$

Table 8. End-to-end fine-tuning settings for FLARE 2022 and BTCV datasets.

Method	Backbone	COVID-19		
		20%	50%	100%
Rand. init.	3D ResNet	73.55±9.33	76.64±7.11	84.73±5.13
MoCo v2 [3]		76.73±9.16	77.94±7.03	85.86±4.92
BYOL [6]		76.69±9.20	77.88±7.15	85.74±5.04
ContrastiveCrop [15]		78.65±8.36	80.61±6.28	87.02±3.11
LoGo [21]		78.60±8.82	80.53±6.75	86.95±3.59
PCRL [23]		79.44±8.44	81.17±6.21	87.31±2.88
PGL [19]		76.77±9.09	78.02±6.93	86.08±4.72
DiRA [7]		78.06±9.04	79.15±6.86	87.43±3.55
Rand. init.	ViT-B	72.80±9.25	76.76±6.90	85.05±4.94
MoCo v3 [4]		77.62±9.17	78.91±6.62	86.32±4.66
DINO [2]		78.49±8.77	80.33±6.28	86.87±4.35
IBOT [24]		79.53±8.05	81.42±5.53	87.55±3.63
SIM [18]		79.85±7.87	81.60±5.05	87.67±2.95
MAE [8]		78.25±8.02	79.78±6.84	86.62±3.27
SemMAE [11]		78.57±7.60	80.47±5.67	86.94±3.44
CMAE [10]		80.05±7.08	81.65±4.92	87.73±3.02
Tang *et al.* [17]		79.59±7.59	81.52±5.05	87.70±3.07
Alice		**83.30±6.04**	**85.23±3.91**	**90.88±1.29**

Table 9. Classification performance of using different pre-training strategies on the COVID19 screening test set. CNN-based SSL methods take the 3D ResNet as their encoder backbone. ViT-based SSL methods take the ViT-B as their encoder backbone. We adopt three label settings (using 20%, 50%, and 100% labeled training data).

nel as the number of classes for prediction. We train the classification model using the AdamW optimizer with a warm-up cosine scheduler of 400 iterations. We use a batch-size of 8 per GPU, an initial learning rate of $5e^{-5}$, a momentum of 0.9 and a decay of $1e^{-5}$ for 10K iterations. We utilize the cross entropy loss. The classification performance is measured by the area under the receiver operator curve (AUC).

B.4. Results

We compare **Alice** with the state-of-the-arts including representative CNN-based SSL methods and ViT-based SSL methods. The results are shown in Tab. 9. **Alice** noticeably surpasses the other state-of-the-art SSL frameworks. We show when using 20% of labeled training data, **Alice** achieves approximately 11% improvement comparing to training from scratch. When employing all labeled training data, the self-supervised pre-training shows 5.83% higher AUC. In practice, the AUC number 85.05 of learning from scratch with entire dataset can be achieved by using pre-trained weights from **Alice** with 50% training data, which indicates that **Alice** can reduce the annotation effort by at least 50% for this task.

Compared with the state-of-the-art CNN-based SSL methods MoCov2, BYOL, PCRL, PGL and DiRA, **Alice** outperforms them at least absolute 3.86% and 3.45% in AUC when using 20% and 100% labeled training data, respectively. Notably, **Alice** achieves much better results than LoGo and ContrastiveCrop, which also design specific strategies to generate semantic-aligned contrastive view pairs. However, these two methods only operate within each image independently and ignore the inter-volume consistency. The superiority performance of **Alice** also reveals the effectiveness of our anatomical semantic alignment strategy.

Compared against strong ViT-based SSL methods, **Alice** significantly outperforms them on all three label settings. The performance gains over the second, third and fourth top-ranked methods are 3.15%, 3.18%, 3.21% and 3.33% on AUC when 100% labels are available. It proves the effectiveness of modeling anatomical invariance and performing semantic alignment to assist the SSL process. Besides, we find that contrastive learning tends to benefit classification task more than MIM, which is consist with many previous studies [10, 18, 12]. Contrastive learning naturally endows the pretained model with strong instance discriminability, while MIM focuses more on learning local relations in input image for fulfilling the reconstruction task [12]. We also notice that the ViT-based methods tend to outperform the CNN-based methods when the number of training data scales up. It reflects that the Vision Transformer is a competitive architecture and the SSL is vital for it to achieve good performance.

C. Ablation Studies

We have conducted additional ablation experiments to further validate the design choices we have made in **Alice**.

C.1. Masking for the target encoder

We investigate whether adopting random masking for the target encoder affects the model performance. As shown

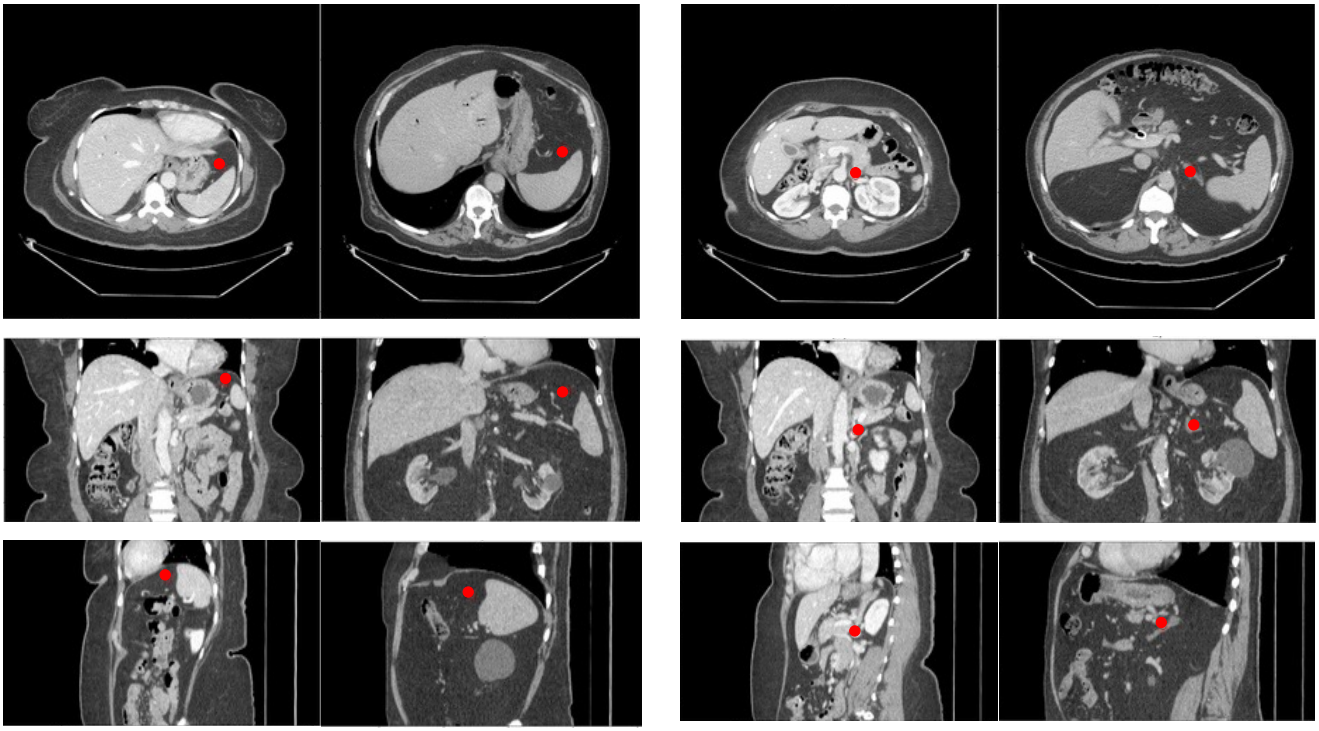

Figure 5. An example of anatomical location matching via SAM [20]. We randomly select an anatomical point in a query CT image, and then use SAM to find its matched point in a key volume from another patient. The red points are selected points in query volume or detected points in key volume.

Masking ratio	DSC on FLARE 2022	DSC on BTCV (offline)
0.75	84.45±2.63	84.30±1.66
0.5	85.22±2.44	85.06±1.35
0.25	86.01±2.27	85.94±1.27
0	**86.87±1.84**	**86.76±0.98**

Table 10. Experiment on whether adopting masking to the input of target encoder. We test different masking ratio settings on FLARE 2022 dataset and BTCV dataset. The segmentation backbone is nnFormer.

in Tab. 10, we observe that using the intact views yields the best results on FLARE 2022 dataset and BTCV dataset. The target encoder provides the online encoder with the contrastive supervision. If target encoder also takes random masked input with degenerated semantic information, the anatomical alignment process will be sub-optimal since the teacher embedding from CASA module may hardly access gloabl information from the original volume crop. Thus, the target encoder in **Alice** uses the whole intact views as inputs.

C.2. Efficacy of combining MIM and CL

We perform experiments on pre-training with different combinations of self-supervised objectives to study the effectiveness of MIM and contrastive learning. Tab. 11 shows the results on FLARE 2022 test set. Overall, employing all objectives achieves best Dice of 86.87%.

Method	MIM ℓ_r	Inter-Volume ℓ_{dv}	Intra-Volume ℓ_{st}	DSC on FLARE 2022
nnFormer baseline	×	×	×	81.33±3.05
	×	✓	×	83.17±2.82
	✓	✓	×	85.66±2.18
	✓	×	✓	85.63±2.11
Alice	✓	✓	✓	**86.87±1.84**

Table 11. Ablation study of different combinations of self-supervised objectives in **Alice** on FLARE 2022 benchmark. The segmentation backbone is nnFormer.

D. More explanations on the feature alignment module

The feature alignment module (CASA) is the one contribution of our paper. We have also thought about using traditional image registration methods (e.g. deformable many-one registration) or some unsupervised learning-based medical image registration methods [1, 22] for alignment. However, a large masking ratio would already erase many image contents and make the masked view quite distinct from the intact one. We found existing medical image registration methods can not work well to solve this problem. Driven

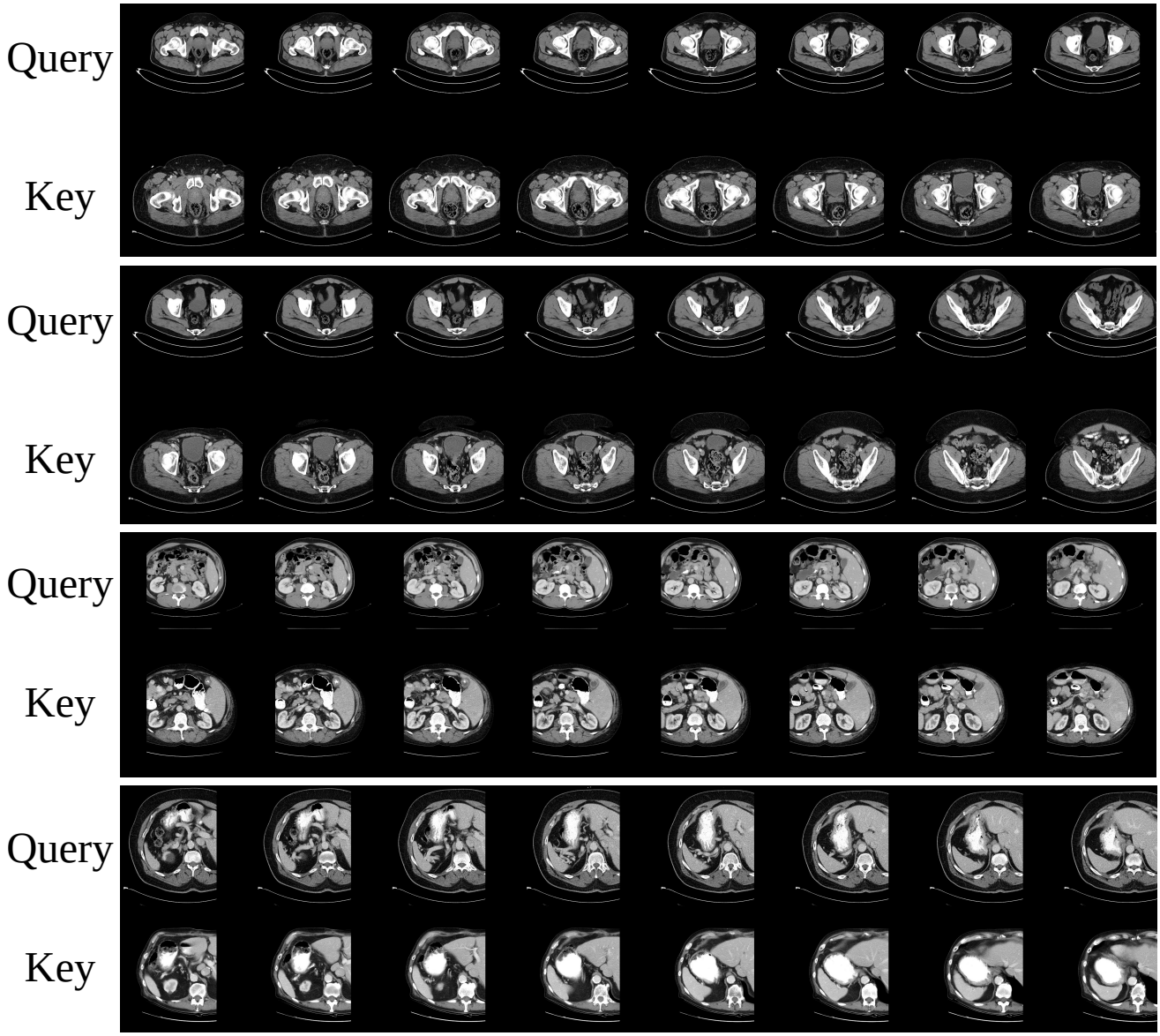

Figure 6. Random anatomy matching results of SAM [20]. We show different query and key crops.

by this limitation, we propose the CASA module to perform alignment in the feature space. "anatomical semantic alignment" is not performing a registration task. So the feature alignment module (CASA) in our **Alice** is not trying to match masked images to full images. It aims to extract aligned features (most correlated) from masked views and augmented views. This module is essentially a self-attention based feature extractor, not a registration algorithm. We compared our module with SIM [18] and CMAE [10] which adopt a specific decoder to generate aligned features. Our method outperforms these two methods in all tasks.

E. BTCV Quantitative Comparisons

In this section, we provide the quantitative comparisons on BTCV offline test set. Note that the ground truth labels of online test set are not accessible. As shown in Fig. 7, **Alice** successfully identifies all organs with high accuracy while it is easy to see that Swin UNETR and nnU-Net produce some under-segmentation and over-segmentation errors. Moreover, as can be seen from the comparison results in the last row of Fig. 7, Swin UNETR and nnU-Net misclassify the spleen (red) as liver (pink) while **Alice** makes the right organ classification. Such superiority of **Alice** owes to the effectiveness of modeling anatomical invariance.

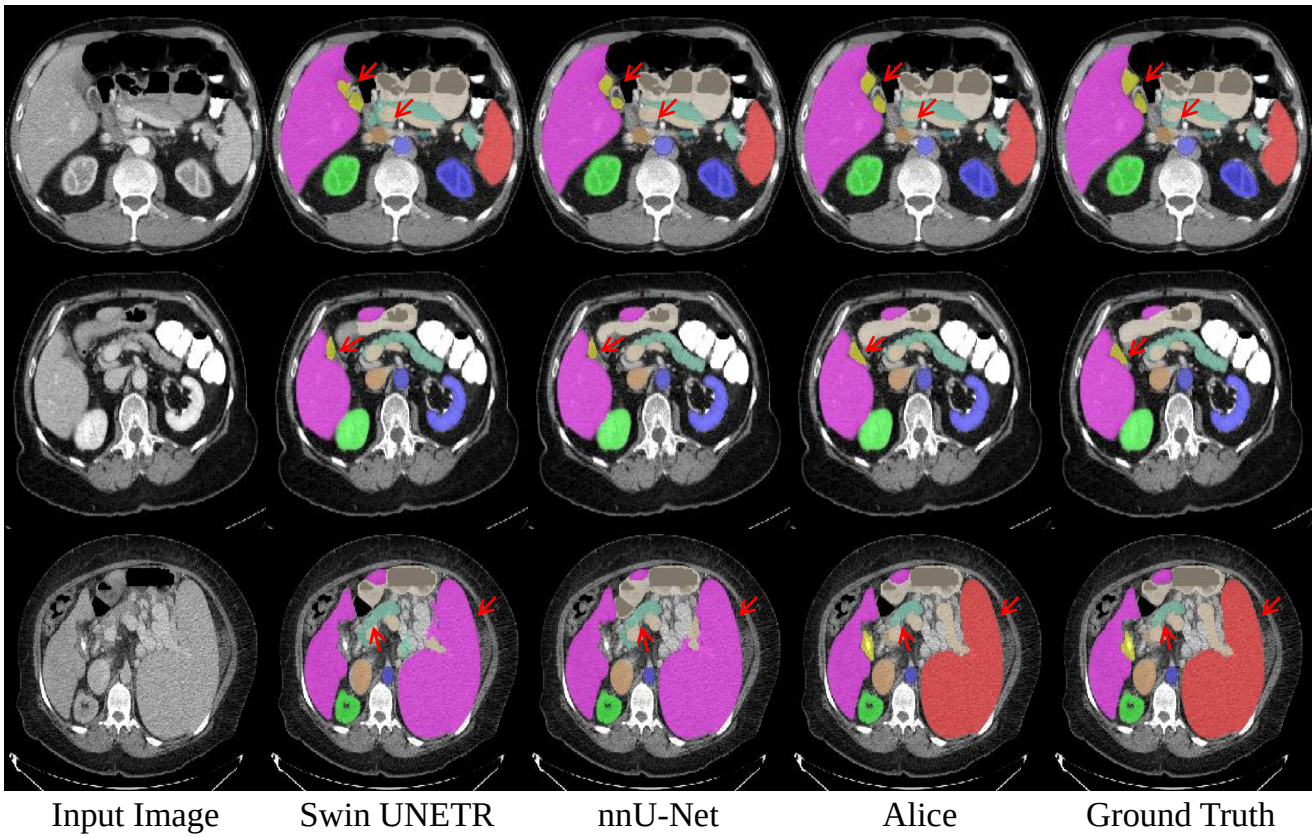

Figure 7. Qualitative visualizations on BTCV offline test set. We compare **Alice** with state-off-the-art segmentation methods, namely Swin UNETR and nnU-Net. The segmentation backbone for **Alice** is nnFormer.

References

[1] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018. 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021. 2

[5] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013. 1

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020. 2

[7] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *CVPR*, pages 20824–20834, 2022. 2

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2

[9] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021. 1

[10] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *arXiv preprint arXiv:2207.13532*, 2022. 2, 4

[11] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided

masking for learning masked autoencoders. *arXiv preprint arXiv:2206.10207*, 2022. 2

[12] Siyuan Li, Di Wu, Fang Wu, Zelin Zang, Baigui Sun, Hao Li, Xuansong Xie, Stan Li, et al. Architecture-agnostic masked image modeling–from vit back to cnn. *arXiv preprint arXiv:2205.13943*, 2022. 2

[13] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE TPAMI*, 2021. 1

[14] Sergey P Morozov, AE Andreychenko, NA Pavlov, AV Vladzymyrskyy, NV Ledikhova, VA Gombolevskiy, Ivan A Blokhin, PB Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020. 1

[15] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *CVPR*, pages 16031–16040, 2022. 2

[16] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 1

[17] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *CVPR*, pages 20730–20740, 2022. 2

[18] Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022. 2, 4

[19] Yutong Xie, Jianpeng Zhang, Zehui Liao, Yong Xia, and Chunhua Shen. Pgl: prior-guided local self-supervised learning for 3d medical image segmentation. *arXiv preprint arXiv:2011.12640*, 2020. 2

[20] Ke Yan, Jinzheng Cai, Dakai Jin, Shun Miao, Dazhou Guo, Adam P Harrison, Youbao Tang, Jing Xiao, Jingjing Lu, and Le Lu. Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. *IEEE Trans. Medical Imaging*, 2022. 1, 3, 4

[21] Tong Zhang, Congpei Qiu, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Leverage your local and global representations: A new self-supervised learning strategy. In *CVPR*, pages 16580–16589, 2022. 2

[22] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10600–10610, 2019. 3

[23] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *ICCV*, pages 3499–3509, 2021. 2

[24] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2