J. Vis. Commun. Image R. 78 (2021) 103145

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

Rethinking pre-training on medical imaging ☆

Yang Wen [a], Leiting Chen [a,b], Yu Deng [c], Chuan Zhou [a,*]

[a] Key Laboratory of Digital Media Technology of Sichuan Province, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan Province, 611731, China

[b] Institute of Electronic and Information Engineering in Guangdong, University of Electronic Science and Technology of China, Chengdu, Sichuan Province, 611731, China

[c] Department of Biomedical Engineering, King's College London, London, UK

ARTICLE INFO

Keywords:
Transfer learning
Medical image analysis
Convolutional neural network
Survival prediction

ABSTRACT

Transfer learning from natural image datasets, such as ImageNet, is common for applying deep learning to medical imaging. However, the modalities of natural and medical images differ considerably, and the reason for the latest medical research preferring ImageNet to medical data is questionable. In this study, we investigated the properties of medical pre-training and its transfer effectiveness on various medical tasks. Through an intuitive convolution-based analysis, we determined the modality characteristics of images. Surprisingly, medical pre-training showed exceptional performance for a classification task but not for a segmentation task since medical data are visually homogeneous and lack morphological information. Using data with diverse modalities helped overcome such drawbacks, resulting in medical pre-training achieving performance comparable to pre-training with ImageNet with considerably fewer samples than ImageNet for both aforementioned tasks. Finally, a study of learned representations and realistic scenarios indicated that while ImageNet is the best choice for medical imaging, medical pre-training has significant potential.

1. Introduction

With the continued development of deep learning and convolutional neural networks (CNNs), transfer learning based on large-scale datasets (*e.g.*, ImageNet [1]) has been employed in industrial applications and research, especially in the medical field where weight transfer of CNNs is important for optimal performance as medical dataset usually contains only a limited number of samples.

In recent years, such methods have been routinely used in CNN-based medical imaging studies and applications, such as respiratory disease classification [2,3]; early-stage skin cancer [4,5], acute intracranial hemorrhage [6], and musculoskeletal abnormality detection [7]; cell segmentation [8]; and early glaucoma diagnosis [9].

Despite the popularity of transfer learning in medical imaging, there has been little study on its precise effects. Moreover, while recent works on natural images have revealed major shortcomings, such as overestimated generalization ability [10] and poor transfer effectiveness even between similar tasks [11], they have not focused on medical imaging, and their findings remain questionable since medical and natural images differ considerably.

Intuitively, visual information of natural and medical images differ significantly. In Fig. 1, in each dataset, compared with medical images, natural images appear diverse and possess more contour details and more colors, reflecting rich visual information. By contrast, the medical images look almost identical, indicating considerably lesser visual information. Accordingly, natural image tasks are usually accomplished by identifying major morphological characteristics (*e.g.*, edges, colors, or shapes) of the primary subjects, while in medical applications, pathologies are identified by detecting small abnormalities and local texture variations, such as bleeding [6,12] and inconsistent structures [7]. CNNs pre-trained with different visual information can not only understand image morphological characteristics in different ways (or acquire different types of morphological awareness) but also learn drastically different types or amounts of morphological awareness from natural and medical images and ultimately facilitate the transfer learning of medical tasks.

Another major difference between natural and medical images lies in their modality, which depends on the camera type and imaging methods used. Recent studies have shown that large modality differences can significantly degrade transfer learning performance [9,13]. Considerable modality differences can be perceived between medical and natural images, and even among medical images.

Given the modality difference between medical and natural images and the existence of different methods for obtaining morphological

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail address: zhouchuan@uestc.edu.cn (C. Zhou).

https://doi.org/10.1016/j.jvcir.2021.103145
Received 21 December 2020; Received in revised form 3 April 2021; Accepted 1 May 2021
Available online 4 May 2021

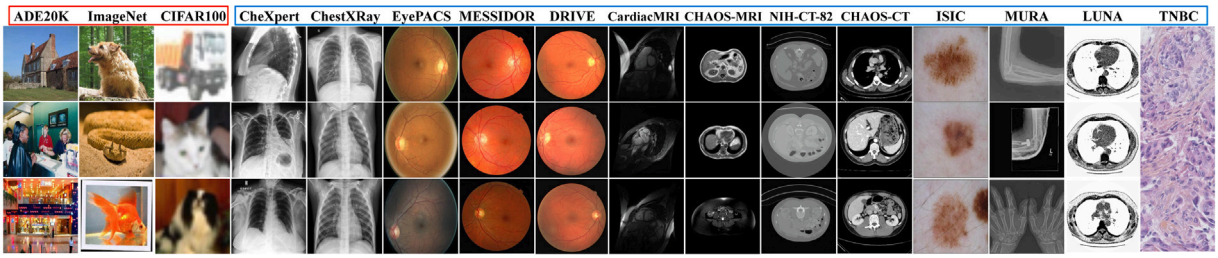


Fig. 1. Example images from natural (below the red box) and medical (below the blue box) image datasets. The medical images are visually homogeneous and lack morphological characteristics.

awareness, it is questionable why recent studies have relied heavily on pre-training with natural image datasets rather than medical images, since similar images intuitively provide close visual information and even relevant hints during transfer learning. Although recent efforts have been made to study transfer learning on medical imaging, they either still pre-training with ImageNet [14] or without thoroughly analyzing the precise properties [15–17]. Furthermore, they have studied only a single classification or segmentation task, and never on survival prediction task, which is not sufficiently comprehensive.

In this work, we performed a fine-grained study of medical-imaging-based transfer learning and evaluated it on several medical tasks. The main results obtained are as follows:

[1] To assess modality differences among images, we propose convolution-based visualization and distance as novel measuring tools. Qualitative and quantitative analyses showed that images with similar visual appearances not only look similar to the naked eye but are also close to each other in the convolved subspace, with a small distance value. Modality similarity can be qualitatively and quantitatively assessed.

[2] We used medical images (instead of ImageNet) to initialize the network, and subsequently fine-tuned and evaluated the network's performance for various medical datasets. We found that while pre-training a model on medical images offered transfer benefits for disease classification, owing to lack of visual information in the medical images, the model cannot effectively learn morphological characteristics, thereby exhibiting poor morphological awareness and poor performance for segmentation tasks. These findings are independent of the number of pre-training samples.

[3] We overcame the above lack of morphological awareness by employing medical data with diverse modalities to provide sufficient morphological and modality information to the model. An analysis of per-layer weight similarity and overall performance demonstrated the effectiveness of this strategy.

[4] Finally, we compared medical and ImageNet pre-training in terms of filter representation and performance for different data regimes. While ImageNet pre-training is superior for medical research, with the availability of an increasing amount of medical data, medical pre-training has considerable potential.

2. Experimental setup

2.1. Datasets and model

We considered a variety of medical imaging datasets with diverse modalities, including prominent X-ray (ChestXRay [2] and CheXpert [3]), fundoscopic imaging (EyePACS [31], MESSIDOR [29], DRISHTI-GS [26], *etc.*), computed tomography (CHAOS-CT [19] and NIH-CT-82 [20]), and magnetic resonance imaging datasets (CardiacMRI [21] and CHAOS-MRI [19]). Datasets without official division were randomly divided into 80%/20% for training/testing. We excluded the target class shared between pre-training and test datasets to prevent task-specific information leakage. For example, in CheXpert and ChestXRay, we excluded the class *pneumonia*, which is shared by both datasets. In total, we selected 20 datasets, of which 17 for medical imaging and 3 for natural images, to conduct our study. Detail information is demonstrated in Table 1.

For most experiments, we report the results of model training from scratch (Random) or with the full ImageNet dataset (ImageNet*). Every experiment was repeated thrice and the metrics of micro-averaged area under curve (AUC) and mean intersection over union (mIoU) were obtained in the form of mean with standard deviation. For convenience, pre-training on medical image datasets is termed *medical pre-training*. To the best of our knowledge, this is the first study to use a dataset collection of such scale and variety for studying the transfer learning in medical fields.

Since a large model is not advantageous for medical tasks [14], we chose the smallest member (*i.e.*, ResNet-18) of the ResNet family [5,13,17,36] as the basic backbone network of our experimental model. We also present results obtained with other classification models with state-of-the-art CNN-based architectures, such as the lightweight model ShuffleNet V2 [37] and the cumbersome model DenseNet-121 [38]. Apart from the commonly studied classification task, we performed experiments on a segmentation task by using the popular segmentation model U-Net [8] and other latest segmentation architectures such as DeepLabv3+ [39], CE-Net [40], HyNet [41], and ET-Net [42]. On the basis of the convolved visual features of the backbone network, U-Net uses an additional oversampling path in parallel to the encoding filters to restore spatial sizes and generate final segmentation outputs. Other segmentation models follow the same basic U-Net structure but with multiple enhancements. DeepLabv3+ and CE-Net enable atrous convolution filters and pyramid pooling modules to enlarge receptive fields and better segment objects at various scales. HyNet combines atrous convolutional filters and parallel up-sampling paths to deeply fuse cross-level visual features and reduce dilution artifacts, and ET-Net contains an extra boundary detection subnetwork to improve segmentation performance on medical objects with blurry boundaries. All segmentation models had the same backbone network and pre-trained parameters for a classification task. The only difference was the extra up-sampling decoder, which required fine-tuning.

2.2. Data preprocessing and augmentation

Since the datasets we used contains images with various sizes and color, some preprocessing procedures are crafted to generate proper inputs for training. Firstly, we crop the useless parts of the images and resize them into 256 × 256 by bilinear interpolation. Secondly, we adopt the CLAHE algorithm [43] to balance the brightness and enhance the details of all images. Besides, due to the limited number of samples, all training images are augmented to expand the datasets, including random rotating (90, 180 and 270 degrees) and random flipping (horizontally, vertically and diagonally), which helps overcome the overfitting problem. ROI extraction is adopted for certain tasks (*e.g.*, optic disk/cup segmentation) similar to that in [9].

Table 1
Statistics of the datasets. Here CLS denotes classification, SEG denotes segmentation, SP denotes survival prediction. For datasets with official division of training and testing sets, the number of images are shown in (number of training images) / (number of testing images), otherwise the total amounts of images.

Name	Modality	Frame Size	# of Images	# of Classes	Task
CheXpert [3]	X-ray	390 × 320	224,316/624	2	CLS
ChestXRay 2017 [2]	X-ray	1000 × 700	5,232/624	2	CLS
LUNA [18]	CT	512 × 512	267	2	SEG
CHAOS-CT [19]	CT	512 × 512	2874	2	SEG
CHAOS-MRI [19]	MRI	320 × 320	992	5	SEG
NIH-CT-82 [20]	CT	512 × 512	7,141	2	SEG
CardiacMRI [21]	MRI	256 × 256	399	3	SEG
ISIC2019 [22–24]	Dermoscopy	1024 × 768	25,331	9	CLS
TNBC [25]	H/E Stained	512 × 512	50	2	SEG
DRISHTI-GS [26]	Fundoscopic	2045 × 1752	50/51	2	SEG
REFUGE [27]	Fundoscopic	1634 × 1634	400/400	2	SEG/CLS
HRF [28]	Fundoscopic	3504 × 2336	45	3	SEG
DRIVE [28]	Fundoscopic	565 × 584	20/20	2	SEG
MESSIDOR [29]	Fundoscopic	2240 × 1488	1,200	4/4	CLS
STARE [30]	Fundoscopic	700 × 650	397	15	SEG/CLS
EyePACS [31]	Fundoscopic	4000 × 4000	35,126	5	CLS
OCT 2017 [2]	OCT	1000 × 300	108,309/1,000	4	CLS
TCGA-GBM [32,33]	H/E Stained	512 × 512	1,523	–	SP
TCGA-LGG [32,33]	H/E Stained	512 × 512	1,380	–	SP
CIFAR100 [34]	Natural Image	32 × 32	50,000/10,000	100	CLS
ImageNet [1]	Natural Image	500 × 375	1.3M/60,000	1000	CLS
ADE20K [35]	Natural Image	1000 × 1000	20,210/2,000	2,603/826	SEG/CLS

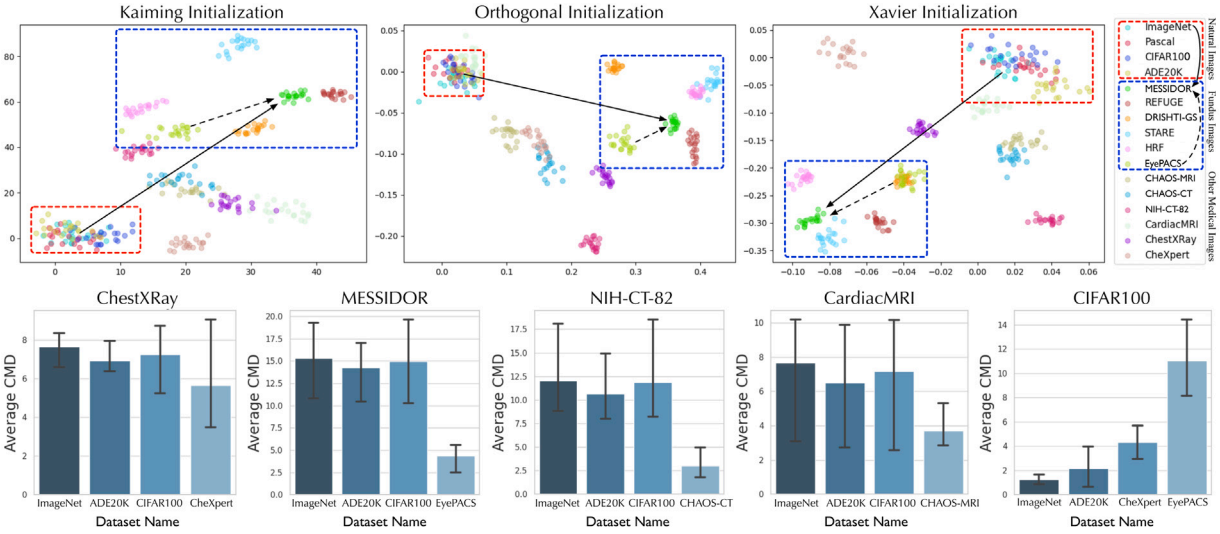


Fig. 2. The *cmv* (top) and *average cmd* (bottom) of natural and medical image datasets. In *cmv*, each dataset was aggregated separately, and the natural (red box) and fundus image datasets (blue box) tended to aggregate, indicating that a CNN could detect the visual similarity of images. Medical imaging datasets with similar modalities show the smallest *average cmd* among all other datasets.

2.3. Experimental details

We implement our network in Python 3.7 based on PyTorch 1.3 platform. All experiments are conducted on Ubuntu 16.04 system with two graphics processing cards, including an NVIDIA GeForce GTX 1080 Ti and an NVIDIA GeForce RTX 2080 Ti, with 22 GB memory in total. We adopt Adam [44] as optimizer with a fixed learning rate of 1e-4, a mini-batch size of 16 and other configuration in default. Early-stopping is triggered when metrics stop improving for 20 iterations. For both classification and segmentation tasks, we use Balanced Cross Entropy as loss functions, where the weighting factors are defined according to different datasets.

3. Convolution-based modality visualization and distance

Previous studies have judged modality differences from subjective image observations or performance degradation [9,13]. Measurements such as H-divergence [45] or maximum mean discrepancy [46], which have been widely used to assess the distance between two data distributions, either require training a task-specific mapping function or are infeasible for a qualitative demonstration. Here, we propose a simple and intuitive technique called *convolution-based modality visualization (cmv)*, which innovatively uses an untrained CNN model to aggregate image features and reduce them into two-dimensional coordinates, to visually evaluate image characteristics. The coordinates are obtained from the following formula:

$$coordinates(S) = \left\{ \frac{1}{n}\sum \mathcal{F}\left(s_1\right), \ldots, \frac{1}{n}\sum \mathcal{F}\left(s_t\right) \right\}, \left\{ s_1, \ldots, s_t \right\} \in S \quad (1)$$

where S is an image dataset, $\mathcal{F}$ is a CNN model (without nonlinear activation, normalization, or dropout), s_t is a subset sampled from S, n is the number of images in s, and t is the total number of times we sampled from S. We present the convolved coordinates of various natural and medical image datasets in Fig. 2. Furthermore, we defined a *convolution-based modality distance (cmd)*, based on the coordinates

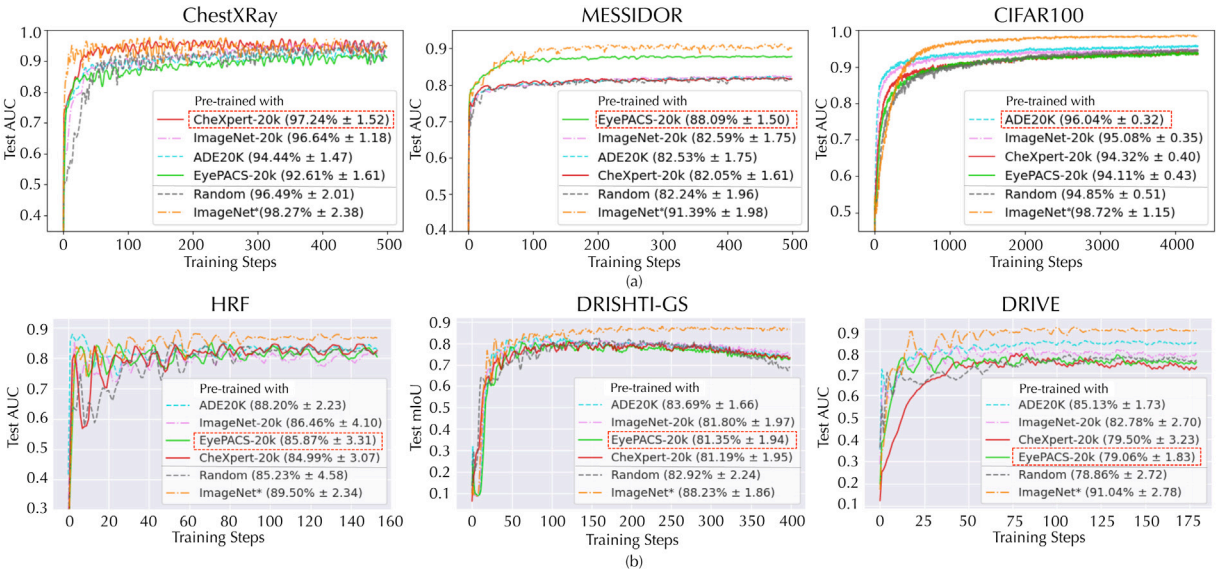


Fig. 3. Results of ResNet-18 pre-training on classification datasets and of (a) classification and (b) segmentation tests. Models pre-trained with classification datasets with the most similar modality (red box) achieved the best AUC for both medical and natural image classification tasks. For segmentation datasets, however, the natural dataset ADE20K brought the best performance, indicating a shortage of morphological features in the medical datasets.

obtained from images, for quantitative evaluation:

$$cmd\left(S_a, S_b\right) = \frac{1}{n^2}\sum_{i=0}^{n}\sum_{j=0}^{n}\sqrt{\left(x_{a_i} - x_{b_j}\right)^2 + \left(y_{a_i} - y_{b_j}\right)^2}$$
$$, \left\{\left(x_k, y_k\right)\right\} = coordinate\left(S_k\right) \tag{2}$$

where S_a and S_b are two image datasets, n is the number of coordinates, and (x, y) are the values of each coordinate. The *cmd* is the mean Euclidean distance of all coordinates between the two datasets. Since the scales of the CNN outputs differ with the initialization type, we used an *average cmd*, defined as follows, as the final measurement:

$$average\ cmd\left(S_a, S_b\right) = \frac{1}{n}\sum_{i=1}^{n}\epsilon_{init_n}\cdot cmd\left(S_a, S_b\right)_{init_n}$$
$$, init = \left\{I_1, \ldots, I_i, \ldots, I_n\right\} \tag{3}$$

where I is the initialization method and ϵ is a factor used to rescale the *cmd* to the same magnitude for each initialization. We chose three popular initialization methods – Kaiming [47], orthogonal [48], and Xavier [49] – for demonstration and made interesting observations.

Modality Similarity Can be Detected by Both Naked Eyes and CNN. In Fig. 2, the *cmv* shows that every dataset aggregated separately, implying that a certain image dataset possessed a specific modality characteristic to cause the images to have different locations in the subspace. Notably, both natural and medical datasets (*e.g.*, fundoscopic imaging data) clearly show aggregation, indicating that the similarity among images can be detected not only by human eyes but also by a CNN. The aggregation of the coordinates of datasets with similar modalities in the *cmv* indicates that a CNN can yield similar features within its intermediate layers. Therefore, since large *cmd* indicates large distance in *cmv* subspace and images with different modalities have a large *cmd*, the features differ significantly when the CNN processes data with different modalities. This suggests that images with similar modalities facilitate easier identification and optimization by CNNs during transfer learning.

Rescaling Factor for Convolution-based Modality Distance. We use the intra convolution-based modality distance (cmd) of ImageNet as ϵ factors to rescale the *cmd* to the same magnitude for each initialization. Specifically, ϵ is 4.4976 for Kaiming [47], 0.0317 for Orthogonal [48] and 0.0162 for Xavier [49].

4. Pre-training on medical imaging data

From the observations in Section 1, it is unclear whether pre-training with medical images with high *modality similarity* (or low *cmd*) improves the CNN performance for medical tasks. The impact of *morphological awareness* on the CNN performance is also unclear. This section discusses the extensive experiments performed to explore the properties of medical pre-training.

4.1. Pre-training on classification datasets

In the first experiment, we pre-trained on four large-scale classification datasets: two natural (ImageNet and ADE-20K [35]) and two medical image datasets (CheXpert and EyePACS). Later, we fine-tuned and tested on classification datasets of ChestXRay, MESSIDOR, and CIFAR100 [34]. As evident in Fig. 2, the modalities of ChestXRay, EyePACS, and CIFAR100 were the closest to those of CheXpert, MESSIDOR, and ADE20K/ImageNet, respectively. For convenience, we denote the model trained with a dataset by the dataset's name in italics (*e.g.*, *EyePACS* represents the model pre-trained on the EyePACS dataset). To eliminate the effect of the number of samples, we used subsets of twenty thousand images, which were named ImageNet-20k, EyePACS-20k, CheXpert-20k, and ADE20K.

Medical Pre-training on Image Datasets with Identical Modalities Outperforms That on Image Datasets with Different Modalities. As Fig. 3a shows, medical pre-training on a dataset with the highest modality similarity was superior to that on other datasets. For ChestXRay, *CheXpert* showed the highest convergence speed and the best AUC (97.24%). For MESSIDOR, *EyePACS* showed significant improvement compared with the other models and achieved a performance level very close to that of *ImageNet**. Notably, *EyePACS* used less than 1/1000 of the number of samples in *ImageNet**. The combination of the observation on CIFAR100 with the preceding information in this paragraph reveals that pre-training on datasets with the highest modality similarity yields the best transfer benefits and features for classification tasks. Similar conclusions can be drawn from observations of other classification networks (ShuffleNet V2 and DenseNet-121) in Table 2,

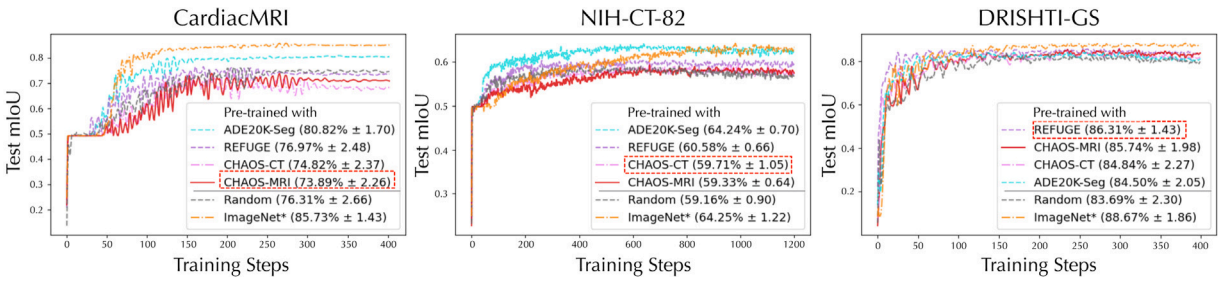


Fig. 4. Results of models pre-trained and tested on a segmentation task. While medical pre-training (red box) leads to high performance of the models on a close target, it results in poor performance if transferred to a completely different target.

Table 2
Results of ShuffleNet V2 and DenseNet-121 initialized by either pre-training on classification datasets or self-supervised learning, and of the classification and segmentation test. The best results are highlighted in **bold**.

Model	Initialization	ChestXRay [2]	MESSIDOR [29]	CIFAR100 [34]	HRF [28]	DRISHTI-GS [26]	DRIVE [50]
ShuffleNet V2 [37]	Random	92.17 ± 2.03	79.66 ± 2.01	94.01 ± 0.68	83.32 ± 5.48	82.81 ± 2.21	79.21 ± 2.93
	CheXpert-20k	**95.42 ± 1.98**[a]	76.06 ± 1.49	93.15 ± 0.42	81.84 ± 3.70	81.24 ± 2.10	79.03 ± 2.41
	EyePACS-20k	90.23 ± 1.84	**87.37 ± 1.54**[a]	93.21 ± 0.39	82.53 ±3.13[a]	81.32 ±2.03[a]	78.03 ±1.93[a]
	ADE20K	95.05 ± 2.07	81.47 ± 1.79	**94.73 ± 0.49**[a]	**85.12 ± 3.29**	**83.12 ± 1.93**	**83.25 ± 2.76**
	SS-Jigsaw [51][b]	92.79 ± 2.53	79.31 ± 2.10	94.25 ± 0.45	80.31 ± 5.36	80.23 ± 1.95	79.19 ± 3.63
	SS-Context [52][b]	93.02 ± 2.34	78.24 ± 1.97	94.16 ± 0.99	80.40 ± 5.61	80.62 ± 1.92	78.56 ± 4.19
DenseNet-121 [38]	Random	97.14 ± 1.94	85.83 ± 1.80	95.59 ± 0.53	86.11 ± 5.29	86.02 ± 1.83	79.35 ± 1.62
	CheXpert-20k	**98.32 ± 1.87**[a]	81.21 ± 1.92	95.43 ± 0.41	84.49 ± 4.41	84.37 ± 1.93	81.34 ± 3.19
	EyePACS-20k	95.28 ± 1.52	**91.63 ± 1.47**[a]	95.30 ± 0.49	86.03 ±3.42[a]	84.72 ±2.01[a]	80.22 ±1.64[a]
	ADE20K	97.91 ± 1.96	89.24 ± 1.87	**97.47 ± 0.31**[a]	**86.41 ± 4.05**	**86.93 ± 1.69**	**86.19 ± 2.03**
	SS-Jigsaw [51][b]	96.42 ± 1.92	81.43 ± 2.01	95.72 ± 0.36	82.83 ± 5.87	81.61 ± 1.83	80.10 ± 2.45
	SS-Context [52][b]	97.06 ± 1.88	83.63 ± 1.83	95.68 ± 0.42	83.10 ± 6.16	82.49 ± 1.94	78.61 ± 1.93

[a]The most similar modality was shared between the pre-training and fine-tuning datasets.
[b]Self-supervised learning methods.

and they indicate that such transfer advantages can be achieved for different deep learning architectures.

Medical Pre-training Fails to Provide Sufficient Morphological Awareness. We also evaluated medical pre-training on three fundus segmentation datasets, namely, HRF [28], DRIVE [50] (retinal vessel), and DRISHTI-GS [26] (optic cup). As evident in Fig. 3b, pre-training on the dataset with the highest modality similarity (*i.e.*, *EyePACS*) failed to yield the best results. Rather, pre-training on a dataset with a large modality distance (*i.e.*, *ADE20K*) yielded the best results. Similar observations were obtained for other backbone networks (ShuffleNet V2 and DenseNet-121), and they are presented in Table 2. Unlike classification tasks, a segmentation task requires rich *morphological awareness* for locating and recognizing the target objects [11]. Since medical images possess considerably less morphological information than natural images, a CNN pre-trained on medical images cannot recognize the morphological characteristics of the target and its performance is inevitably poor. The lack of *morphological awareness* appears to affect the performance more strongly than the lack of *modality similarity*. The effect of different tasks or the number of samples should be assessed to determine which of the two factors (*i.e.*, *morphological awareness* and *modality similarity*) is more critical for medical pre-training.

4.2. Pre-training on segmentation datasets

In the second experiment, for the evaluation of medical pre-training (similar to [17]), we replaced the classification datasets with segmentation datasets, which comprised the natural image dataset ADE20K-Seg (segmentation version of ADE20K) and three medical datasets, namely, CHAOS-CT, CHAOS-MRI, and REFUGE-OD (optic disk part of REFUGE [27]). After the pre-training of the models, we fine-tuned the models and determined their performance for three segmentation datasets, namely, CardiacMRI [21], NIH-CT-82 [20], and DRISHTI-GS. Specifically, CHAOS-CT and CHAOS-MRI are subparts of the CHAOS dataset. While CHAOS-CT pertains to the liver, CHAOS-MRI concerns

Table 3
Results for the latest segmentation models pre-trained and tested on a segmentation task. The best results are highlighted in bold.

Model	Initialization	CardiacMRI	NIH-CT-82	DRISHTI-GS
DeeplabV3+ [39]	Random	80.21 ± 2.41	60.21 ± 0.96	83.67 ± 2.02
	CHAOS-CT	79.35 ± 3.59	59.74 ±1.02[a]	83.91 ± 1.79
	CHAOS-MRI	80.05 ±2.38[a]	60.09 ± 0.87	83.73 ± 1.64
	REFUGE	78.52 ± 2.57	60.12 ± 1.03	**85.29 ± 1.64**[a]
	ADE20K-Seg	**81.56 ± 2.93**	**65.27 ± 1.12**	85.14 ± 2.14
	SS-Jigsaw [51][b]	80.77 ± 3.44	61.24 ± 0.75	81.02 ± 1.72
	SS-Context [52][b]	80.95 ± 2.20	60.83 ± 0.88	81.24 ± 1.52
CE-Net [40]	Random	80.34 ± 1.71	61.04 ± 0.98	83.13 ± 2.10
	CHAOS-CT	80.01 ± 2.35	60.31 ±0.68[a]	82.32 ± 1.69
	CHAOS-MRI	79.47 ±1.73[a]	60.42 ± 0.83	83.30 ± 1.79
	REFUGE	80.45 ± 1.85	59.85 ± 0.90	**85.44 ± 1.34**[a]
	ADE20K-Seg	**82.19 ± 2.32**	**65.21 ± 1.01**	85.09 ± 2.01
	SS-Jigsaw [51][b]	80.53 ± 2.42	61.32 ± 0.84	81.31 ± 1.45
	SS-Context [52][b]	80.84 ± 2.10	62.56 ± 0.96	82.24 ± 1.82
HyNet [41]	Random	80.93 ± 2.13	62.42 ± 0.86	83.48 ± 1.41
	CHAOS-CT	80.55 ± 2.49	61.45 ±1.21[a]	82.69 ± 2.13
	CHAOS-MRI	80.83 ±2.04[a]	62.03 ± 0.89	82.51 ± 1.66
	REFUGE	80.09 ± 2.24	61.46 ± 1.13	**86.11 ± 2.10**[a]
	ADE20K-Seg	**83.47 ± 1.92**	**67.88 ± 1.21**	85.47 ± 1.93
	SS-Jigsaw [51][b]	81.22 ± 2.34	62.34 ± 0.87	82.13 ± 1.73
	SS-Context [52][b]	81.09 ± 2.02	62.73 ± 1.19	81.84 ± 2.10
ET-Net [42]	Random	81.75 ± 1.98	62.73 ± 0.95	83.75 ± 1.95
	CHAOS-CT	81.29 ± 2.41	62.12 ±1.02[a]	83.10 ± 2.01
	CHAOS-MRI	80.83 ±1.78[a]	61.49 ± 0.92	82.64 ± 1.34
	REFUGE	81.23 ± 1.92	63.31 ± 1.31	**86.23 ± 1.36**[a]
	ADE20K-Seg	**83.59 ± 2.08**	**67.97 ± 0.96**	85.62 ± 1.94
	SS-Jigsaw [51][b]	81.01 ± 2.30	63.01 ± 0.79	82.34 ± 1.25
	SS-Context [52][b]	80.93 ± 1.96	62.69 ± 0.80	82.81 ± 2.10

[a]The most similar modality was shared between the pre-training and fine-tuning datasets.
[b]Self-supervised learning methods.

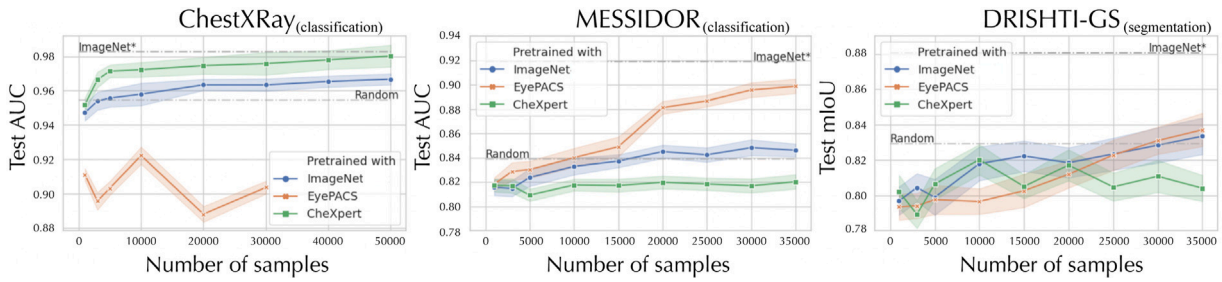


Fig. 5. Results for different numbers of pre-training samples. Medical pre-training constantly outperformed natural image pre-training on classification tasks, but not on a segmentation task.

the liver, kidney, and spleen. The test datasets CardiacMRI, NIH-CT-82, and DRISHTI-GS pertain to the endocardium and epicardium, pancreas, and optic cup, respectively. The highest modality similarities were observed between CardiacMRI and CHAOS-MRI, between NIH-CT-82 and CHAOS-CT, and between REFUGE-OD and DRISHTI-GS. Besides U-Net, we present results for the other latest segmentation models, namely, DeepLabv3+ [39], CE-Net [40], HyNet [41], and ET-Net [42].

Medical Segmentation Pre-training Also Fails to Introduce Morphological Awareness. As Fig. 4 shows, medical pre-training showed worse performance than even *Random* for CardiacMRI and NIH-CT-82, while *ADE20K-Seg* showed the best results for both these datasets. Similar performance degradation for the CardiacMRI and NIH-CT-82 datasets can be observed in the case of other latest segmentation models in Table 3. Evidently, segmentation pre-training can provide satisfactory representation, similar to classification training, and therefore, the aforementioned poor performance of medical pre-training is attributed to the medical images. As mentioned in Section 4.1, without sufficient morphological information, the model cannot acquire sufficient *morphological awareness* from medical images. Although an exception occurs when targets in the pre-training and fine-tuning stages are similar (*e.g.*, the targets, namely, optic disk and optic cup, are located close to each other in REFUGE and DRISHTI-GS), in which case the highest modality similarity leads to the best performance, the absence of *morphological awareness* remains the main reason for the poor performance of the model initialized by medical pre-training on the other two datasets (*i.e.*, CardiacMRI and NIH-CT-82).

4.3. Observation on self-supervised learning methods

Recent studies on self-supervised learning methods have shown the suitability of the methods for initializing CNNs with sufficient visual features and their potential to outperform transfer learning methods. However, to date, no comparison has been made between medical pre-training and their self-supervised learning counterparts. Here, we present results of two popular self-supervised learning methods (SS-Jigsaw [51] and SS-Context [52]) for a comprehensive comparison. We directly pre-trained CNNs using the self-supervised methods with the target datasets considered in this study (*i.e.*, ChestXRay, MESSIDOR, CIFAR100, HRF, DRISHTI-GS, DRIVE, CardiacMRI and NIH-CT-82).

Self-supervised Learning with Medical Images Also Fails to Provide Sufficient Morphological Awareness. As shown in Table 2, for both ShuffleNet V2 and DenseNet-121 backbone networks, while the self-supervised learning models showed better performance on some datasets (*e.g.*, CIFAR100, and DRIVE) compared with *Random*, the models using transfer learning strategies still showed the best performance. Furthermore, for the latest segmentation models (Table 3), self-supervised learning methods could improve their performance and render them somewhat better than *Random* for large datasets (CardiacMRI and NIH-CT-82), but not for small datasets (DRISHTI-GS). These observations indicate that despite some solid results of previous studies [51,52], self-supervised learning is not trivial in the medical domain. Owing to the lack of morphological information in medical images, self-supervised learning methods cannot introduce sufficient morphological awareness into CNNs either. Another possible reason for the poor performance of self-supervised models is that the number of samples in some datasets (*e.g.*, DRISHTI-GS, HRF, and DRIVE) is small and may not be sufficient to properly initialize a CNN model. One piece of evidence for this hypothesis is that the performance of self-supervised models trained with a large number of samples (*e.g.*, CIFAR100, ChestXRay, CardiacMRI, and NIH-CT-82) is generally better than *Random*, while that of self-supervised models trained with small datasets is not.

4.4. Analysis of the number of samples

This section discusses our investigation of whether the number of pre-training samples influenced previous observations that medical data fail to provide sufficient *morphological awareness* to CNN model. Experiments were performed on three datasets: two classification datasets (Chest-X-Ray and MESSIDOR) and a segmentation dataset (DRISHTI-GS).

Medical pre-training Constantly Outperforms ImageNet pre-training on Classification Tasks. The performance gaps for ImageNet and medical pre-training are apparent. As observed in Fig. 5, medical pre-training constantly performed better than *ImageNet* on a classification task. Despite being pre-trained on only fifty thousand samples, *CheXpert* showed performance very close to that of *ImageNet**, which was trained on nearly 1.3 million images. For MESSIDOR too, high performance was observed. This demonstrates a significant *advantage* of medical pre-training: data with higher modality similarity facilitate a significant performance improvement on a classification task.

Increasing Sample Size Does Not Compensate for Lack of Morphological Awareness. For the segmentation of DRI-SHTI-GS, medical pre-training did not significantly improve the model performance when the sample size was increased. Despite its very high *modality similarity*, *EyePACS* constantly performed worse than *Random*. Thus, increasing the number of pre-training samples does not help overcome the lack of *morphological awareness*.

5. Overcoming drawbacks of medical pre-training and further analysis

5.1. Overcoming drawbacks of medical pre-training

Although medical pre-training facilitates effective transfer under certain circumstances, it has a key drawback: **lack of morphological awareness**. The images in a single medical dataset are highly homogeneous and therefore provide very limited morphological information. Furthermore, they may cause the CNN to be confined to a particular

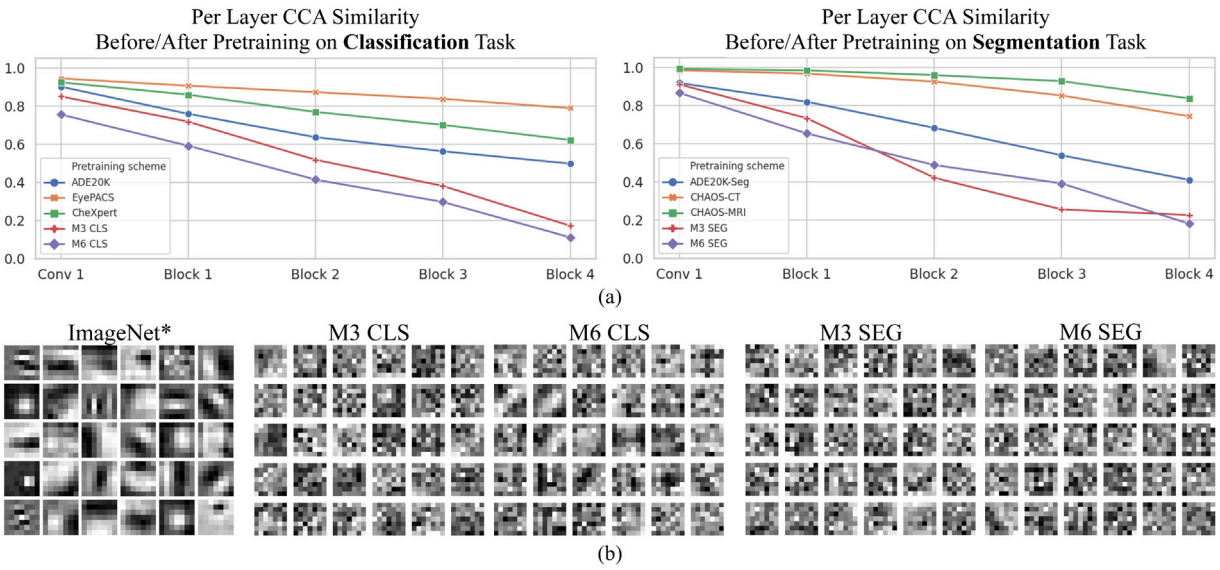


Fig. 6. Per-layer CCA similarities before/after training on different datasets/collections (a) and the corresponding visualization of *conv1* filters (b). With an increase in the modality and variety of the data, larger changes are apparent in the filters in M3 and M6, resulting in richer representations.

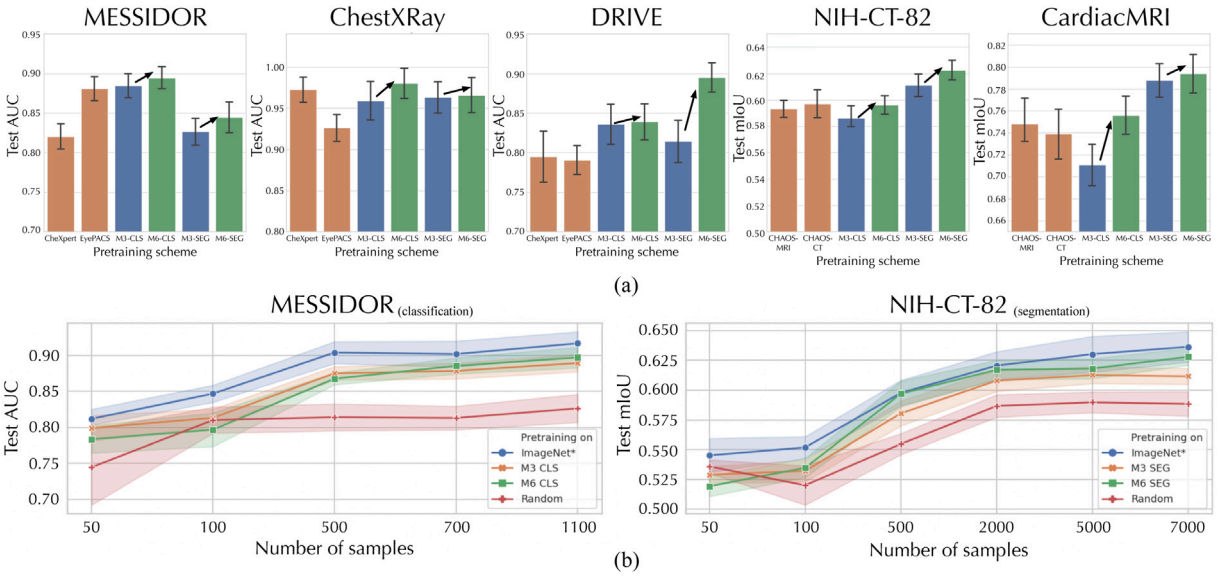


Fig. 7. Performance results for (a) different datasets/collections and (b) different numbers of fine-tuning samples. *M6* outperformed *M3* and the single medical dataset, and despite containing considerably fewer samples than *ImageNet**, it showed performance comparable to that of *ImageNet**.

domain and render its transfer to another medical task difficult. Since the morphological information in a single-modality medical dataset is limited, adding more medical data with diverse modalities might help improve the *morphological awareness*, since more modalities and morphological features can be provided to the CNN. We used several collections of datasets to evaluate the effect of medical data with different types of modalities on pre-training. For classification pre-training, two collections – M3 CLS (CheXpert, EyePACS, and ISIC [22–24]) and M6 CLS (M3 CLS + OCT [2], MURA [7], and STARE-CLS [30]) – were used. For segmentation pre-training too, two collections – M3 SEG (CHAOS-CT, CHAOS-MRI, and TNBC [25]) and M6 SEG (M3 SEG + LUNA [18], REFUGE-OD, and STARE-SEG) – were used.

Representational Analysis of CNN Filters. We first compared the per-layer representational similarity before and after pre-training using Singular Vector Canonical Correlation Analysis (SV-CCA) [53], as shown in Fig. 6a. Clearly, *M3* and *M6* provided more filtering modifications compared to the previous scheme (*i.e.*, ADE20K, EyePACS *etc.*), indicating that more features were captured by the CNN. This plot also suggests that pre-training on medical data collections could introduce comparable features in the natural image dataset (*i.e.*, ADE20K) and that both medical classification and segmentation tasks could facilitate satisfactory network initializations. Finally, a comparison of the similarity between *conv1* and the previous scheme shows that *M3* and *M6* caused more changes in the low-level convolutional layer, indicating that more task-independent features were learned. Additionally, we visualized the *conv1* filter of the CNN in Fig. 6b. It is evident that *M6* brings more Gabors than *M3* and make *conv1* closer to that of *ImageNet**, indicating that the use of large amount of medical data with diverse modalities for pre-training resulted in representational improvement of CNN.

Table 4
Numbers of WSIs, patients, originally extracted patches, and valid patches. Patches extracted from background parts in WSIs have been excluded.

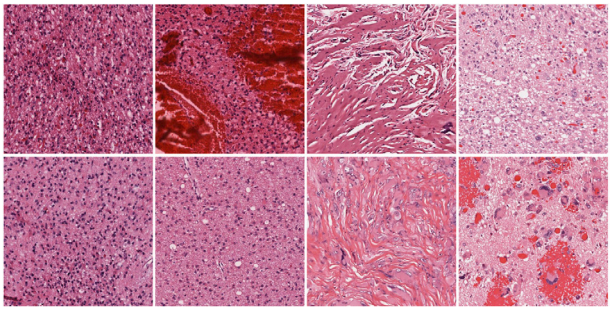
Dataset	TCGA-LGG	TCGA-GBM
Number of patients	387	349
Number of WSIs	1380	1523
Number of patches	275453	339912
Number of valid patches	103512	112565

More Modalities Lead to Better Performance. We evaluated the performance of *M3* and *M6* on five datasets. In Fig. 7a, it is evident that *M6* outperformed *M3* and the previous scheme on all five tasks in several aspects. First, the representational improvement resulting from the use of *M6* led to better performance, which indicated enhanced *morphological awareness*. Second, *M3* and *M6* did not degrade the performance, indicating strong transfer robustness. Third, both *M6 CLS* and *M6 SEG* outperformed their M3 counterparts, indicating that for both segmentation and classification, *morphological awareness* could be improved by involving diverse modalities. Thus, pre-training on medical dataset collections is effective for overcoming the aforementioned drawback of medical pre-training and yields satisfactory results.

5.2. Further analysis: High potential and limitations of medical pre-training

Fig. 7b shows a major advantage and a disadvantage of medical pre-training. On the one hand, compared with natural images, medical images with high *modality similarity* provide more transfer benefits in medical applications. In particular, *medical pre-training* with tens of thousands of samples can lead to performance comparable to that with *ImageNet**, which contains millions of samples. With the increased use of deep learning methods in healthcare over the past year, large amounts of large-scale medical data (*e.g.*, CheXpert and EyePACS) have become available, and they could facilitate the development of a medical version of ImageNet in the near future. In such an event, since considerably more modalities of data can be used compared with our study, the transfer effectiveness of medical pre-training is likely to exceed that of pre-training with ImageNet.

On the other hand, it is noteworthy that *ImageNet** constantly outperformed medical pre-training on both MESSIDOR and NIH-CT-82 for different data regimes, suggesting that ImageNet pre-training is currently the best option for medical research. Although a recent study [11] demonstrated that pre-training with ImageNet is not drastically different from slightly longer training with random weights, but these studies were based on large-scale natural image datasets. As evidenced by our previous experiments in Figs. 3 and 4, *ImageNet* showed the best performance on almost all medical datasets, especially outperforming *Random* on the small-scale datasets, indicating that training with random weights could not lead to the same performance as training with *ImageNet*. The reason why an observation identical to that of the previous study (the performance of the model initialized by pre-training on ImageNet is similar to that of the model trained with random weights) was not made for the medical image datasets in the present study is twofold: First, the number of samples in the natural image datasets (118,287 images in the COCO dataset [54] and 15,000 images in the PASCAL VOC dataset [55]) is significantly greater than that in most of the medical datasets (*e.g.*, 267 images in the LUNA dataset, 45 images in the HRF dataset, and 399 in the CardiacMRI dataset). Second, as mentioned, compared with medical images, natural images and their corresponding annotations are considerably more varied, which allows for more morphological awareness. Owing to these reasons and the collection of sufficient medical data being several years away, ImageNet transfer is currently the most economical and effective option.



(a) TCGA-LGG (b) TCGA-GBM

Fig. 8. Some sample patches from two TCGA datasets. The patches in each dataset are from the same patient.

5.3. Apply medical pre-training on survival prediction

Apart from classification and segmentation tasks, survival analysis is also an important part of computer-aided medical image analysis and plays a critical role in current clinical practice [56–58]. Therefore, we performed medical pre-training on a survival analysis task and assessed the transfer robustness. We focused on brain cancer in our study and used two public cancer survival datasets containing whole slide pathological images (WSIs) with high resolution from The Cancer Genome Atlas (TCGA) [32]. Specifically, we conducted experiments on two subtypes of brain cancer in TCGA projects: lower-grade glioma (LGG) and glioblastoma (GBM). We adopted the annotations of the vital status and overall survival time from a previous study [33] and followed the patch selection procedure used in [56]. All patches were extracted with a size of 512 × 512 pixels from WSIs. The numbers of WSIs and patients in each dataset are presented in Table 4 and sample patches are shown in Fig. 8. In the experiments, the concordance index (C-index) [58] was used as the metric for performance evaluation. It was defined as the quality of rankings and calculated as follows:

$$C\text{-}index = \frac{1}{n} \sum_{i \in \{1 \dots N | \delta_i = 1\}} \sum_{t_j > t_i} I\left[f_i > f_j\right], \tag{4}$$

where n is the number of pairs compared, t is the true observed value, f is the risk factor, and $I[\cdot]$ is an indicator function. The C-index ranges from zero to one, and the higher its value, the better the prediction of the models.

Survival Prediction Also Benefits from Medical Pre-training. To validate the effectiveness of medical pre-training on the survival prediction task, we first used two recent deep learning models (ResNet-18 and DenseNet-121) with two clustering methods (K-means and spectral clustering) to directly predict the survival time. As shown in Table 5, for the ResNet-18 and DenseNet-121 models with either K-means or spectral clustering, the use of the ImageNet dataset for weight initializations yielded the best performance. Nevertheless, compared with random weight initialization, medical pre-training helped improve the survival prediction performance by $\geq 4.04\%$, and even up to 11.39% (compare with *M6-SEG* and *Random* on DenseNet-121 + SP). This indicates transfer benefits resulting from medical pre-training.

Furthermore, we applied medical pre-training to the feature extractors of two recent deep-learning-based survival prediction frameworks, namely WSISA [56] and DeepAttnMISL [58], to further validate the effectiveness of medical pre-training. In particular, WSISA was evaluated in combination with a recent multitask learning method MTLSA [59], which is specifically designed for survival prediction tasks. Apart from deep-learning-based models, we also present results from two previous state-of-the-art methods (RSF [60] and BoostCI [61]). As shown in Table 6, compared with random weight initialization, medical pre-training improved the performance of WSISA + MTLSA by up to 3.84% for TCGA-LGG and by up to 5.65% for TCGA-GBM. For DeepAttnMISL,

Table 5
Survival prediction results (in percentage) for different feature extractors and clustering methods. KM and SP denote K-means clustering and spectral clustering, respectively. The clustering number was set to six.

Model	Initialization	TCGA-LGG	TCGA-GBM
ResNet-18 + KM	Random	47.21 ± 4.81	45.35 ± 4.40
	M6-CLS	52.76 ± 4.42	51.09 ± 5.39
	M6-SEG	53.01 ± 3.93	52.17 ± 4.83
	ImageNet*	**55.51 ± 4.04**	**54.48 ± 4.21**
ResNet-18 + SP	Random	47.87 ± 4.59	40.21 ± 5.86
	M6-CLS	51.91 ± 5.49	48.03 ± 4.62
	M6-SEG	53.18 ± 5.03	49.10 ± 3.49
	ImageNet*	**55.40 ± 4.28**	**54.79 ± 4.01**
DenseNet-121 + KM	Random	48.65 ± 5.13	43.89 ± 6.11
	M6-CLS	56.30 ± 3.92	53.64 ± 4.21
	M6-SEG	53.76 ± 4.08	50.29 ± 4.42
	ImageNet*	**58.38 ± 3.92**	**60.49 ± 4.24**
DenseNet-121 + SP	Random	51.81 ± 5.35	45.81 ± 5.41
	M6-CLS	56.48 ± 3.49	56.32 ± 4.81
	M6-SEG	53.12 ± 4.32	57.20 ± 4.32
	ImageNet*	**57.24 ± 4.37**	**60.19 ± 3.90**

Table 6
Survival prediction results (in percentage) for the latest models, where the feature extractors of deep-learning-based models were initialized in different ways. The best results are highlighted in bold.

Model	Initialization	TCGA-LGG	TCGA-GBM
RSF [60]	–	48.24 ± 4.09	50.53 ± 3.98
BoostCI [61]	–	47.75 ± 5.33	52.77 ± 4.18
WSISA [56] + MTLSA [59][a]	Random	43.45 ± 5.11	42.90 ± 5.66
	M6-CLS	47.29 ± 3.48	48.55 ± 4.41
	M6-SEG	46.03 ± 5.06	45.28 ± 3.93
	ImageNet*	**49.78 ± 4.70**	**53.91 ± 4.72**
DeepAttnMISL [58][a]	Random	45.93 ± 5.71	49.89 ± 4.35
	M6-CLS	51.73 ± 4.66	53.54 ± 4.87
	M6-SEG	54.34 ± 3.94	55.22 ± 3.90
	ImageNet*	**56.81 ± 3.49**	**57.87 ± 3.76**

[a]Deep learning-based models.

medical pre-training yielded an improvement by up to 8.41% for TCGA-LGG and by up to 5.33% for TCGA-GBM, which confirmed the effectiveness of medical pre-training. Nonetheless, the models initialized with ImageNet still performed the best on both datasets, indicating the same conclusion as in Section 5.2: ImageNet transfer is currently the most economical and effective option.

6. Conclusion

We investigated the effectiveness of medical pre-training. Through *cmv* and *cmd*, we determined the modality characteristics of images and showed the effectiveness of medical pre-training on a classification task. We also revealed drawbacks related to the generalization ability and morphological awareness, identified their origin to be the lack of visual variety in medical images, and showed that the drawbacks could be overcome by introducing additional medical data with diverse modalities. Finally, we compared the two pre-training schemes for real-world scenarios and found that pre-training with ImageNet is still the best choice owing to its advanced visual representation and generalization tolerance. However, we also demonstrated that medical pre-training has significant potential.

CRediT authorship contribution statement

Yang Wen: Conception and design, Analysis and interpretation, Data collection, Critical revision, Approval of the manuscript, Agreement to be accountable, Statistical analysis. **Leiting Chen:** Analysis and interpretation, Critical revision, Approval of the manuscript, Agreement to be accountable. **Yu Deng:** Conception and design, Analysis and interpretation, Data collection, Approval of the manuscript, Agreement to be accountable, Statistical analysis. **Chuan Zhou:** Conception and design, Analysis and interpretation, Critical revision, Approval of the manuscript, Agreement to be accountable, Statistical analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by Key-Area Research and Development Program of Guangdong Province, China (No. 2019B010136003), and Sichuan Science and Technology Program, China (No. 2019YJ0176/2019YJ0177/2019YFQ0005).

References

[1] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[2] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131.
[3] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 590–597.
[4] A. Esteva, B. Kuprel, R.A. Novoa, J.M. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.
[5] Z. Yu, X. Jiang, F. Zhou, J. Qin, D. Ni, S. Chen, B. Lei, T. Wang, Melanoma recognition in dermoscopy images via aggregated deep convolutional features, IEEE Trans. Biomed. Eng. 66 (4) (2019) 1006–1016.
[6] H. Lee, S. Yune, M. Mansouri, M. Kim, et al., An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets, Nat. Biomed. Eng. 3 (3) (2019) 173–182.
[7] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R.L. Ball, MURA: Large dataset for abnormality detection in musculoskeletal radiographs, 2017.
[8] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[9] S. Wang, L. Yu, X. Yang, C. Fu, P. Heng, Patch-based output space adversarial learning for joint optic disc and cup segmentation, IEEE Trans. Med. Imaging 38 (11) (2019) 2485–2495.
[10] S. Kornblith, J. Shlens, Q.V. Le, Do better imagenet models transfer better, 2018, arXiv: Computer Vision and Pattern Recognition.
[11] K. He, R. Girshick, P. Dollár, Rethinking imagenet pre-training, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4918–4927.
[12] Y. Wen, L. Chen, L. Qiao, Y. Deng, S. Dai, J. Chen, C. Zhou, Symptom and pathology report generation for ophthalmic diseases in fundus images, in: 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2020, pp. 349–356.
[13] Q. Dou, C. Ouyang, C. Chen, H. Chen, et al., Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss, 2018, pp. 691–697.
[14] M. Raghu, C. Zhang, J. Kleinberg, S. Bengio, Transfusion: Understanding transfer learning for medical imaging, in: Advances in Neural Information Processing Systems, 2019, pp. 3342–3352.
[15] K. Yan, X. Wang, L. Lu, R.M. Summers, DeepLesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations, 2017, arXiv: Computer Vision and Pattern Recognition.
[16] Z. Zhou, V. Sodha, M.R. Siddiquee, R. Feng, N. Tajbakhsh, M.B. Gotway, J. Liang, Models genesis: Generic autodidactic models for 3D medical image analysis, 2019, pp. 384–393.
[17] S. Chen, K. Ma, Y. Zheng, Med3D: Transfer learning for 3D medical image analysis, 2019, arXiv: Computer Vision and Pattern Recognition.
[18] A.A.A. Setio, A. Traverso, T. De Bel, M.S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M.E. Fantacci, B. Geurts, 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 (2017) 1–13.

[19] A.E. Kavur, N.S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D.D. Pham, S. Chatterjee, P. Ernst, S. Özkan, et al., CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation, Med. Image Anal. 69 (2021) 101950.
[20] H.R. Roth, L. Lu, A. Farag, H.-C. Shin, et al., Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 556–564.
[21] A. Andreopoulos, J.K. Tsotsos, Efficient and generalizable statistical models of shape and appearance for analysis of cardiac MRI, Med. Image Anal. 12 (3) (2008) 335–357.
[22] N.C.F. Codella, D.A. Gutman, M.E. Celebi, B. Helba, Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging, hosted by the international skin imaging collaboration (ISIC), 2016, arXiv: Computer Vision and Pattern Recognition.
[23] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (1) (2018) 180161.
[24] M. Combalia, N.C.F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, A.C. Halpern, S. Puig, J. Malvehy, BCN20000: dermoscopic lesions in the wild, 2019, arXiv: Image and Video Processing.
[25] P. Naylor, M. Laé, F. Reyal, T. Walter, Segmentation of nuclei in histopathology images by deep regression of the distance map, IEEE Trans. Med. Imaging 38 (2) (2018) 448–459.
[26] J. Sivaswamy, S.R. Krishnadas, G.D. Joshi, M. Jain, A.U.S. Tabish, Drishti-GS: Retinal image dataset for optic nerve head(ONH) segmentation, in: IEEE International Symposium on Biomedical Imaging, 2014.
[27] J.I. Orlando, H. Fu, J.B. Breda, K. van Keer, et al., REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs, Med. Image Anal. 59 (2020) 101570.
[28] A. Budai, R. Bock, A. Maier, J. Hornegger, G. Michelson, Robust vessel segmentation in fundus images, Int. J. Biomed. Imaging 2013 (2013).
[29] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, et al., Feedback on a publicly distributed image database: the Messidor database, Image Anal. Stereol. 33 (3) (2014) 231–234.
[30] A.D. Hoover, V. Kouznetsova, M. Goldbaum, Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, IEEE Trans. Med. Imaging 19 (3) (2000) 203–210.
[31] EyePacs, 2015, By California Healthcare Foundation, available at https://www.kaggle.com/c/diabetic-retinopathy-detection/. (Accessed 15 July 2020).
[32] C. Kandoth, M.D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J.F. McMichael, M.A. Wyczalkowski, et al., Mutational landscape and significance across 12 major cancer types, Nature 502 (7471) (2013) 333–339.
[33] P. Mobadersany, S. Yousefi, M. Amgad, D.A. Gutman, J.S. Barnholtz-Sloan, J.E.V. Vega, D.J. Brat, L.A. Cooper, Predicting cancer outcomes from histology and genomics using convolutional networks, Proc. Natl. Acad. Sci. 115 (13) (2018) E2970–E2979.
[34] A. Krizhevsky, G. Hinton, et al., Learning Multiple Layers of Features from Tiny Images, Citeseer, 2009.
[35] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ade20k dataset, 2016, arXiv preprint arXiv:1608.05442.
[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016, pp. 770–778.
[37] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 116–131.
[38] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[39] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 801–818.
[40] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu, CE-Net: Context encoder network for 2D medical image segmentation, IEEE Trans. Med. Imaging (2019).
[41] Y. Wen, L. Chen, L. Qiao, C. Zhou, S. Xi, R. Guo, Y. Deng, An efficient weakly-supervised learning method for optic disc segmentation, in: 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2020, pp. 835–842.
[42] Z. Zhang, H. Fu, H. Dai, J. Shen, Y. Pang, L. Shao, Et-net: A generic edge-attention guidance network for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 442–450.
[43] J. Skilling, S.F. Gull, Algorithms and applications, Lecture Notes in Comput. Sci. 14 (3) (2010) 83–132.
[44] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, Comput. Sci. (2014).
[45] S. Bendavid, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J.W. Vaughan, A theory of learning from different domains, Mach. Learn. 79 (1) (2010) 151–175.
[46] K.M. Borgwardt, A. Gretton, M.J. Rasch, H. Kriegel, B. Scholkopf, A.J. Smola, Integrating structured biological data by Kernel Maximum Mean Discrepancy, Bioinformatics 22 (14) (2006) 49–57.
[47] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
[48] A.M. Saxe, J.L. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, 2013, arXiv preprint arXiv:1312.6120.
[49] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.
[50] J. Staal, M.D. Abràmoff, M. Niemeijer, M.A. Viergever, B. Van Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE Trans. Med. Imaging 23 (4) (2004) 501–509.
[51] M. Noroozi, P. Favaro, Unsupervised learning of visual representations by solving jigsaw puzzles, in: European Conference on Computer Vision, Springer, 2016, pp. 69–84.
[52] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, D. Rueckert, Self-supervised learning for medical image analysis using image context restoration, Med. Image Anal. 58 (2019) 101539.
[53] M. Raghu, J. Gilmer, J. Yosinski, J. Sohldickstein, SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability, 2017, pp. 6076–6085.
[54] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[55] M. Everingham, S.A. Eslami, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, Int. J. Comput. Vis. 111 (1) (2015) 98–136.
[56] X. Zhu, J. Yao, F. Zhu, J. Huang, Wsisa: Making survival prediction from whole slide histopathological images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7234–7242.
[57] E. Wulczyn, D.F. Steiner, Z. Xu, A. Sadhwani, H. Wang, I. Flament-Auvigne, C.H. Mermel, P.-H.C. Chen, Y. Liu, M.C. Stumpe, Deep learning-based survival prediction for multiple cancer types using histopathology images, PLoS One 15 (6) (2020) e0233678.
[58] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, J. Huang, Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks, Med. Image Anal. 65 (2020) 101789.
[59] Y. Li, J. Wang, J. Ye, C.K. Reddy, A multi-task learning formulation for survival analysis, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1715–1724.
[60] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, et al., Random survival forests, Ann. Appl. Stat. 2 (3) (2008) 841–860.
[61] A. Mayr, M. Schmid, Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations, PLoS One 9 (1) (2014) e84483.