A Knowledge-based Learning Framework for Self-supervised Pre-training Towards Enhanced Recognition of Medical Images

Wei Chen, Chen Li, Dan Chen, and Xin Luo

***Abstract*—Self-supervised pre-training has become the priory choice to establish reliable models for automated recognition of massive medical images, which are routinely annotation-free, without semantics, and without guarantee of quality. Note that this paradigm is still at its infancy and limited by closely related open issues: 1) how to learn robust representations in an unsupervised manner from unlabelled medical images of low diversity in samples? and 2) how to obtain the most significant representations demanded by a high-quality segmentation? Aiming at these issues, this study proposes a knowledge-based learning framework towards enhanced recognition of medical images, which works in three phases by synergizing contrastive learning and generative learning models: 1) Sample Space Diversification: Reconstructive proxy tasks have been enabled to embed *a priori* knowledge with context highlighted to diversify the expanded sample space; 2) Enhanced Representation Learning: Informative noise-contrastive estimation loss regularizes the encoder to enhance representation learning of annotation-free images; 3) Correlated Optimization: Optimization operations in pre-training the encoder and the decoder have been correlated via image restoration from proxy tasks, targeting the need for semantic segmentation. Extensive experiments have been performed on various public medical image datasets (e.g., CheXpert and DRIVE) against the state-of-the-art counterparts (e.g., SimCLR and MoCo), and results demonstrate that: The proposed framework statistically excels in self-supervised benchmarks, achieving 2.08, 1.23, 1.12, 0.76 and 1.38 percentage points improvements over SimCLR in AUC/Dice. The proposed framework achieves label-efficient semi-supervised learning, e.g., reducing the annotation cost by up to 99% in pathological classification.**

***Index Terms*—Self-supervised Pre-training, Medical Images, Classification, Segmentation, Generative Learning, Contrastive Learning.**

This work was supported by the National Key Research and Development Program of China (No. 2018YFB0204301) and the Natural Science Foundation of Hunan Province of China (No. 2022JJ30666).

Wei Chen is currently a professor with the College of Computer Science, National University of Defense Technology, Changsha 410073, China (e-mail:chenwei@nudt.edu.cn).

Chen Li and Xin Luo are with the College of Computer Science, National University of Defense Technology, Changsha 410073, China (e-mail:lichen14@nudt.edu.cn, luoxin13@nudt.edu.cn).

Dan Chen was a HEFCE research fellow with the University of Birmingham, United Kingdom. He is currently a professor with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: dan.chen@whu.edu.cn).

Wei Chen and Chen Li contributed equally to this work. Corresponding author: Chen Li.

I. Introduction

THE success of nowadays biomedical research and clinical practices have largely relied on automated recognition of massive medical images, sustaining fine-grained interpretation of the physiological and pathological states of organs, tissues, and lesions. These medical images are routinely annotation-free, without semantics, and without guarantee of quality. It still remains an active research area to reach reliable conclusions based on the results of critical tasks such as classification and segmentation of such medical images, as routine end-to-end models for image recognition largely rely on excessive labeling by human experts.

Self-supervised learning has proved powerful in learning representations without the need for large labelled datasets [1]. Self-supervised pre-training has become essential in harsh scenarios like medical image recognition to obtain state-of-the-art performance using unlabelled data [2]. Self-supervised pre-training routinely aims to pre-train an Auto-encoders (AE) model on a large amount of unlabelled medical images. It then adopts the well-trained model [1] to downstream tasks with "optimal" initialization ensured, possibly complemented with alternative models when necessary.

Cutting-edge methods of self-supervised pre-training targeting on image recognition are largely established on *contrastive learning* [3], which centers on how to "learn to compare" to construct a high-quality representation space. Contrastive learning methods can be *context-instance contrast* and *instance-instance contrast*:

- Context-instance contrast models the mutual information (MI) between the local feature and its global context, and the representation space may be optimized by maximizing the MI. These approaches (e.g., InfoMax [4]) can extract the most discriminative local representations for downstream classification tasks. Note that MI measurement is highly computing-intensive, and performance bottleneck needs to be tackled in this context.
- The latter directly measures the similarity between different samples. It then extracts the instance-level representations by pulling the positive (similar) pairs together and pushing the negative(dissimilar) pairs apart. These approaches (e.g., SimCLR [5] and MoCo [6]) become dominant in classification tasks with performance competitive with supervised-based alternatives. However, the

[1]All mentions of the "model" in this paper refer to Auto-encoder.

performance degrades when recognizing medical images due to insufficient sample diversity.

Unfortunately, contrastive learning methods generally cannot suffice in dense prediction tasks, and the segmentation of medical images is exactly the case. These tasks demand correlated optimization between the encoder and decoder, where contrastive learning is designed to optimize encoders only, leaving decoder training unattended.

Generative learning is another self-supervised paradigm that learns the context-instance representations by restoring the original data distribution from transformations. Generative learning does not assume downstream tasks in advance, which can then provide fast and consistent initialization for classification and segmentation tasks. Note that its performance is limited compared to contrastive learning in most tasks.

Self-supervised pre-training is still at its infancy in image recognition despite the success that has been achieved. When handling killer applications in scenarios as harsh as medical image recognition targeting at clinical practices, this paradigm is refrained by the closely related open issues:

(1) How to learn robust representations in an unsupervised manner from unlabelled medical images of low diversity in samples? Unlabelled medical images are routinely with only insignificant inter-class differences. Insufficient sample diversity is a constant under this circumstance, while sufficient positive/negative pairs are mandatory for any successful contrastive learning.

(2) How to obtain the most significant representations demanded by a high-quality segmentation? Segmentation as a dense prediction task demands collaboration between the encoder and the decoder. Contrastive learning is encoder-oriented only, while the performance of solutions based on generative learning is not satisfied.

This study first needs to extend the sample space. a priori knowledge of target tissues in medical images of different modalities may help in enriching the stylistic and structural diversity of the sample space, while direct brute-force learning of the unlabelled images with insignificant differences does not apply. After that, contrastive learning may construct more diverse positive/negative pairs to extract instance-level representations, while generative learning is capable of learning context-level representations. Note that generative learning excels in the co-initialization of the autoencoder. It is desirable to bridge generative and contrastive learning to co-optimize the encoder and decoder. Thus high-quality representations may be obtained for segmentation tasks by befitting from both methods' merits.

Aiming at these issues, this study proposes a knowledge-based learning framework (**TOWER**) **TOW**ards **E**nhanced **R**ecognition of medical images, which works in three phases:

- Sample Space Diversification (Section III-B): Reconstructive proxy tasks have been designed to perform the nonlinear translation and random masked reconstruction based on a priori knowledge from clinic practices. Transformed images are then obtained via these tasks with the style and structural diversity of sample space enriched.
- Enhanced Representation Learning (Section III-C): The transformed images form the basis of constructing positive/negative sample pairs of higher diversity, which sustains the need for contrastive learning. Informative noise contrast estimation (InfoNCE) loss regularizes the feature space extracted by the encoder, which can enhance the representation learning of annotation-free images.
- Correlated Optimization (Section III-D): Generative learning makes full use of the powerful representations from the last phase and applies MSE loss to guide the optimization of the encoder-decoder, which reconstructs the transformed images and enhances the representation learning towards style and structural context. The correlated optimization then bridges contrastive and generative learning and serves the need for semantic segmentation.

Note that the proposed method defines the regions of interest (ROI) specified into the shape of different masks (Section III-B) characterizing typical medical images, e.g., physiological measures of blood vessels related to the optic disc of retinal images. The pre-trained model can extract the target tissue features in the ROI by masking these regions and reconstructing them afterwards, i.e., making use of the important a priori knowledge:

- For retinal images, the rays-wise mask applies to reconstruct the physiological relationships between the optic disc and blood vessels;
- For X-ray images, the stripe-wise mask applies to reconstruct the texture information of the bones distributed in the stripe region;
- For CT images, the block-wise mask applies to reconstruct the distribution of abdominal organs.

Extensive experiments have been performed on various public medical image datasets (e.g., CheXpert and DRIVE) against the state-of-the-art counterparts (e.g., SimCLR and MoCo). The proposed framework's performance (AUC/Dice) and convergence have been evaluated. TOWER's label efficiency has also been examined with different percentages of partially labelled images in a semi-supervised manner.

The main contributions of this study are as follows:

- This study develops a knowledge-based learning framework to recognize medical images without annotations via self-supervised pre-training. The framework significantly improves downstream tasks' performance with enhanced convergence and label efficiency.
- To the best of our knowledge, the proposed framework is the first to tackle the problem of insufficient diversity of contrastive learning for the recognition of medical images.
- A correlated optimization between encoder and decoder is proposed to provide significant representations for initializing decoder demanded by high-quality segmentation of medical images.

II. Related Work

Recognition of unlabelled images had attracted tremendous attentions in the machine learning community, and it remained an intriguing issue to learn robust representations without annotations. Studies undertaken for this purpose centering on pre-training Auto-encoders generally followed two directions:

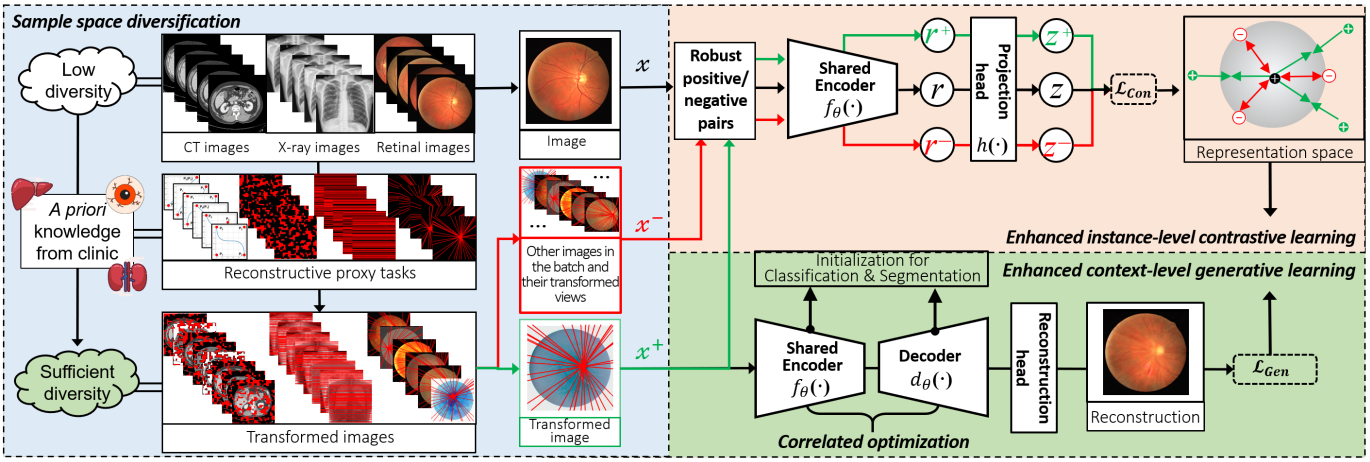


Fig. 1. Overall design of TOWER framework. $\boldsymbol{r}$ and $\boldsymbol{z}$ respectively denote *representations* and *embeddings*. The green and red arrows respectively represent *positive* and *negative pairs*.

1) optimizing the encoder via contrastive learning, and/or 2) co-optimizing the encoder and the decoder via generative learning. The most salient works along these directions were introduced as follows.

Chen et al. developed a simple yet effective *Contrastive Learning* framework SimCLR v1 [5] to optimize the encoder. SimCLR v1 explored data enhancement strategies on two symmetric encoder-mlp branches for contrastive learning. SimCLR v1 achieved comparable performance to supervised-based in natural image classification tasks with a large batch size (8196) required to ensure data diversity.

He et al. adopted a similar idea in their method of MoCo v1 [6] with a momentum encoder. The method introduced a dynamic dictionary to store negative samples with no need for large batch size. MoCo v1 saved at least 48% GPU memory and 22% runtime with a similar performance obtained on V100 GPUs [7] (instead of TPUs in [5]).

Mathilde et al. proposed a cluster-based contrastive learning method (SwAV) [8]. SwAV made comparisons between features and different prototypes via clustering [9] instead of relying on feature pairs. SwAV was superior to existing work in natural image classification but not the case for dense prediction tasks such as medical image recognition.

To initialize the encoder and the decoder at the same time, diverse proxy tasks had been proposed to aid generative learning for more effective pre-training via [10]–[12]. Attempts had then made along the direction of *Generative Learning*:

Zhou et al. proposed a unified framework (Model Genesis [13]) that integrated various proxy tasks to transform images with encoder-decoder initialization, including nonlinear transformation, local shuffling, and in/out painting. By predicting the original images from the transformation, the framework enabled self-supervised representation learning for CT/MRI 3D image analysis. Note that generative learning methods generally were not able to compete with supervised pre-training counterparts when handling medical images [14].

He et al. proposed Masked AutoEncoder (MAE) inspired by the success of masked-language modeling [15]. MAE adopted block-wise masks in model training to reconstruct the randomly-masked input images, thus the structural semantics in the target images might be better captured. Block-wise masking strategies might not directly apply to medical image analysis as task-related semantic characteristics were not guaranteed.

Inspired by the successes of the existing work, this study aimed at self-supervised pre-training towards enhanced recognition of medical images via the synergy of contrastive learning and generative learning: 1) to enrich the diversity of medical images with a priori knowledge, 2) to enhance self-supervised learning in terms of instance-level and context-level representations, and 3) to provide high-quality initialization for dense prediction tasks.

III. Knowledge-based Learning Framework Towards Enhanced Recognition of Medical Images

This section first presents the overall design of TOWER. This section then details the working mechanism of TOWER in three aspects: 1) sample space diversification, 2) enhanced representation learning, and 3) correlated optimization.

A. Overall Design

Fig.1 gives an overview of TOWER framework, where a 2D U-Net [16] is selected as the encoder-decoder with parameters θ, denoted as $f_\theta(\cdot)$ and $d_\theta(\cdot)$. The backbone of the encoder $f_\theta(\cdot)$ is a ResNet-50 [17]-based network with an MLP-based classification head $h(\cdot)$. The AE model receives an input $X \in \mathbb{R}^{N\times H\times W\times C}$, which is a randomly sampled batch with N images; the output is the dense prediction $Y \in \mathbb{R}^{N\times H\times W\times C}$ with the same resolution as X. The objective of model training is to properly initialize both the encoder and decoder to serve the need of downstream tasks with high-quality representations.

As a self-supervised learning framework, TOWER aims to optimize the θ from unlabelled medical images X so that $f_\theta(\cdot)$ and $d_\theta(\cdot)$ can be efficiently fine-tuned using only a few labelled examples when transferring to downstream tasks, e.g., classification and segmentation. TOWER operates as follows:

- Reconstructive proxy tasks $\phi(\cdot)$ embed a priori knowledge from clinic practices. Knowledge-based nonlinear translation (Section III-B1) and masked reconstruction

(Section III-B2) are proposed to enrich the stylistic and structural diversity of sample space, respectively;
- Contrastive learning routine (based on SimCLR v1 [5]) constructs positive/negative sample pairs from the diversified sample space (Section III-C1). The InfoNCE [18] loss function is used to regularize the encoder to learn instance-level representations.
- Generative learning (enhanced Model Genesis [13] in this study) restores the transformed images from reconstructive proxy tasks (Section III-C2). The MSE loss function is used to regularize the encoder and decoder to make consistent reconstructions and learn context-level representations.
- Contrastive and generative learning share the auto-encoder and the strengths between the two are complementary (Section III-D). Contrastive learning optimizes the encoder and provides more powerful representations for generative learning to restore stylistic and structural context. Generative learning optimizes the encoder and decoder, providing significant representations demanded by a high-quality segmentation for contrastive learning.

This design assumes that embedding a priori knowledge can offer an unrivalled opportunity for unsupervised representation learning of medical images. Contrastive learning and generative learning are bridged via reconstructive proxy tasks, and the two are then mutually enhanced for instance-level and context-level unsupervised representation. Correlated optimization of both the encoder and decoder becomes possible with fine-tuned initialization to sustain tasks of medical image classification and dense prediction.

B. Sample Space Diversification: a priori knowledge-embedded reconstructive proxy tasks

1) Enriching stylistic diversity via knowledge-based non-linear translation: In order to augment the medical images, a nonlinear translation proxy task has been designed. This task operates centering on nonlinear transformation, which can change pixel-wise values in an array according to a specific nonlinear mapping relationship. The design here aims to utilize its merit in extending the solution space of linear problems into a non-linear variant [19].

Medical images can be characterized by their special imaging mechanism, i.e., different intensity values in most medical images convey various implicit semantics. Such a priori knowledge has been embedded in the nonlinear translation proxy task. Taking CT images for example, different organs and tissues absorb X-rays differently, and thus doctors can roughly discriminate different regions via their CT values (Hounsfield Unit, HU). The HU values of livers are mostly in the range of [45,65], kidneys in the range of [20,40], and lung parenchyma in the range of [-850,-910]. The HU values of water are 0, and the HU values of air are -1000.

Generally speaking from the perspective of a medical image, changing its pixel values and transforming the overall style will alter the semantic mapping relationships. In this sense, nonlinear translation holds potentials in enriching the stylistic diversity of the original sample space. TOWER designs

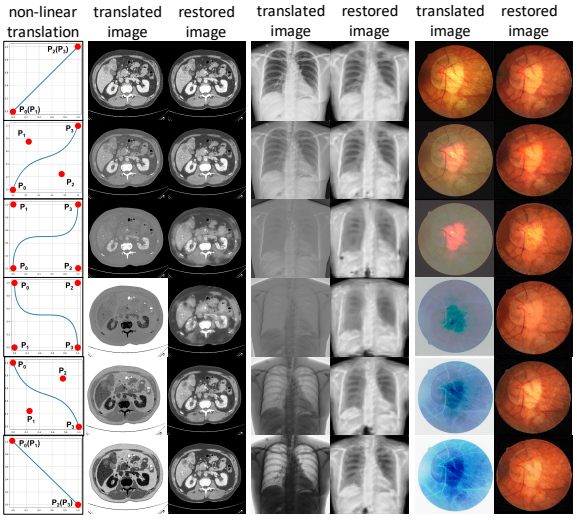


Fig. 2. Illustrations of the nonlinear translations proxy task: six sets of Bézier curves-based functions (1st col) followed by translated images and corresponding reconstructed results.

multiple sets of monotonic invertible functions, which allow the values of each pixel to be restored after changing under given rules. In other words, this design enables invertible transformations of the image style. Bézier Curve[2] is applied to generating the above functions:

$$\text{Bézier}(P, n, t) = \sum_{i=0}^{n} \binom{n}{i} (1-t)^{n-i} \cdot t^i \cdot P_i, \qquad (1)$$

where P denotes the set of interpolation points $\{P_i|_{i=1}^{n}\}$, and t is an independent variable in the range [0,1]. The Bézier curve then forms by interpolating the endpoints and the control points.

TOWER implements the nonlinear translation proxy task upon the cubic Bézier (n=3) as follows:

$$\begin{aligned} p' &= \text{Bézier}(\{P_0, P_1, P_2, P_3\}, 3, p) \\ &= P_0(1-p)^3 + 3P_1(1-p)^2 p + 3P_2(1-p)p^2 + P_3 p^3, \end{aligned} \qquad (2)$$

where p denotes the pixel-wise value in the normalized x_n, p' is the transformed value in the translated x'_n. P_0, P_3 are endpoints and P_1, P_2 are control points.

Fig.2 illustrates the translation functions: (1) they increase monotonically when $P_0 = (0,0)$ and $P_3 = (1,1)$ (shown in the 1st, 2nd and 3rd rows), (2) they decrease monotonically when $P_0 = (0,1)$ and $P_3 = (1,0)$ (shown in the 4th, 5th and 6th rows). Note that the translation functions are linear (shown in the 1st and 6th rows) when $P_0 = P_1$ and $P_2 = P_3$.

2) Enriching structural diversity via knowledge-based masked reconstruction: Inspired by MAE [15], TOWER introduces randomly masked reconstruction as a proxy task to augment the translated images x'_n. A priori knowledge from the clinic is embedded into masks in this course, which considers the structure of common target tissues in medical images of different modalities. This design aims to enrich the structural diversity of the sample space.

The regions of interest (ROI) are specified into the shape of different masks (see Fig.3 for the masks and the masked

[2]https://pomax.github.io/bezierinfo

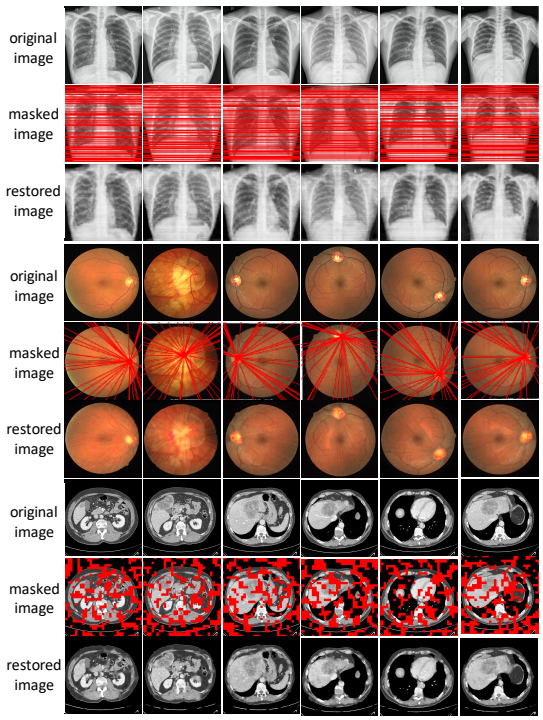


Fig. 3. Illustration of the randomly masked reconstruction proxy task for three modalities: X-ray, retinal fundoscopic, and CT images from top to bottom. Stripe-wise, raw-wise, and block-wise matrices apply to mask the original images. TOWER reconstructs the structural information by predicting the pixel values for each masked pixel. TOWER can then restore the masked images and learn the structural context.

images): (1) for retinal images, the rays-wise mask reconstructs the physiological relationships between the optic disc and blood vessels, (2) for X-ray images, the stripe-wise mask reconstructs the texture information of the bones distributed in the stripe region, and (3) for CT images, the block-wise mask reconstructs the distribution of abdominal organs distributed in the block region.

Taking retinal images for example, the optic disc and blood vessels are the two most common tissues. They are the two most important ROIs in downstream tasks:

(1) The blood vessels of normal eyeballs are emitted from the optic disc. Analogizing the optic disc to a starting point, the vessels can be regarded as rays emanating from the starting point. The starting point of the ray-wise mask is located in the brightness point in the retina images. A rays-wise mask can reasonably simulate the above physiological relationships.

(2) In addition, the diameter of the optic disc of normal people is about 1.5mm, while the diameter of the optic cup is approximately 1/3 of the optic disc, about 0.5mm. Moreover, the average diameter of blood vessels with uneven thickness is about 0.1mm. The mask can then be set according to the rays' thickness and the starting point's diameter.

After obtaining the mask $Mask$, the reconstruction proxy task obtains the transformed images $x_n'' = x_n' \times Mask$ via multiplying x_n' with $Mask$ pixel by pixel. TOWER extracts the context of the target tissue in the ROI by masking these regions and reconstructing them. The two sets of training schemes, i.e., (1) random mask reconstructions and (2) nonlinear translation, are used in a hybrid manner in this study (Fig.4).

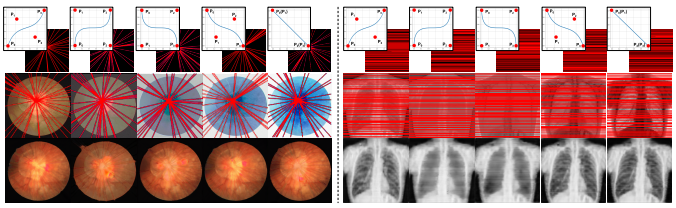


Fig. 4. Reconstructive proxy task integrating a hybrid training scheme: The first rows represent the hybrid schemes; The second rows represent the transformed images; and the last rows represent the reconstruction. TOWER can

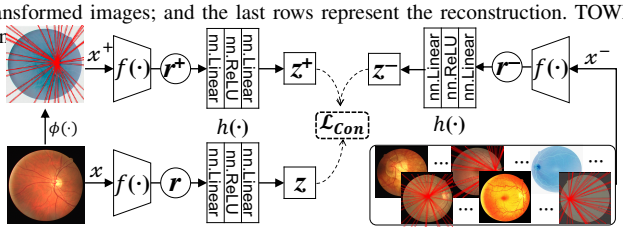


Fig. 5. The contrastive learning workflow. One sample and its transformed views by reconstructive proxy tasks are defined as positive pairs. The remaining samples in this batch and their transformed views are negative pairs.

C. Enhanced unsupervised representation learning

Given the image batch $X = \{x_1, x_2, ..., x_n, ..., x_N\}$ and its transformed views $X'' = \{x_1'', x_2'', ..., x_n'', ..., x_N''\}$, TOWER bridges the contrastive and generative learning via reconstructive proxy tasks. It then enhances the instance-level and context-level unsupervised representation learning towards recognition of medical images in two complementary aspects.

1) Instance-level representation learning via contrastive learning: TOWER customizes the contrastive learning (SimCLR [5]) workflow on the diversified sample space as shown in Fig.5. TOWER defines the sample x_n and its transformed views x_n'' as positive pairs: $x_n{}^+$. The remaining samples and their transformed views are the negative pairs of x_n: x_n^-.

For the sample and its positive/negative pairs ($x_n/x_n{}^+/x_n^-$), a base encoder $\boldsymbol{f}_\theta(\cdot)$ extracts the representations ($\boldsymbol{r_n}/\boldsymbol{r_n^+}/\boldsymbol{r_n^-}$). An MLP as the projection head $\boldsymbol{h}(\cdot)$ maps the representations to embeddings ($\boldsymbol{z_n}/\boldsymbol{z_n^+}/\boldsymbol{z_n^-}$).

The cosine similarity function $sim(a, b)$ then measures the similarity between pairwise embeddings via the dot product between ℓ_2 normalized a and b:

$$sim(a, b) = a^\intercal b/(\|a\| \, \|b\|), \tag{3}$$

Finally, InfoNCE [18] as the contrastive loss function $\mathcal{L}_{Con}$ applies to (1) pull the representations of positive pairs together and (2) push the representations of negative pairs apart, defined as follows:

$$\mathcal{L}_{Con} = \frac{-1}{N}\sum_{n=1}^{N} \log \frac{e^{\left[sim\left(z_n, z_n^+\right)/\tau\right]}}{e^{\left[sim\left(z_n, z_n^+\right)/\tau\right]} + \sum_{z_n^-} e^{\left[sim\left(z_n, z_n^-\right)/\tau\right]}}, \tag{4}$$

where $\tau > 0$ is a scalar temperature (set as 0.1 [5]).

The representation space of the encoder can then be optimized by minimizing the distances of $(\boldsymbol{z_n}, \boldsymbol{z_n^+})$ and maximizing the distances of $(\boldsymbol{z_n}, \boldsymbol{z_n^-})$.

2) Context-level representation learning via generative learning: TOWER enhances the generative learning (Model Genesis [13]) workflow by restoring the transformed images

from more diverse transformations, i.e., nonlinear translation and random masking (Section III-B).

U-Net (an encoder-decoder architecture [16]) makes the dense prediction $y_n = d_\theta(f_\theta(x_n''))$ based on the transformed images X''. The masked regions of X'' evolves invisibly in the training process. The control points set $\{P_i|_{i=1}^n\}$ also evolves with different Bézier Curves generated for nonlinear translation.

Note that such a hybrid transforming scheme ensures that the X'' can not be reconstructed by fitting an interpolation function. Reconstructing medical images from these transformations aids learning context-level representations. For example, reconstructing the optic disc and blood vessels from partially masked retinal images contributes to learning the local context of these tissues. Restoring the correct HU values from style-translated CT images contributes to learning the global context of whole images.

Generative learning restores the transformed image by optimizing the following loss function.

$$\mathcal{L}_{Gen} = \sum_{n=1}^{N} \ell_{mse}(x_n, y_n), \tag{5}$$

where ℓ_{mse} is the mean squared error (MSE) function; The objective is 1) to keep y_n the same as the original image x_n, and 2) to ensure the encoder-decoder model learning the context representations.

D. Correlated optimization between the encoder and decoder

Contrastive learning and generative learning routines excel respectively in extracting instance-level and context-level representations. It is desirable to complement the two with each other to enhance the unsupervised representation learning process.

Contrastive learning specializes in optimizing the encoder, which may provide more powerful representations for generative learning to restore stylistic and structural context. Generative learning's merit in optimizing the encoder and decoder can provide representations significant enough for contrastive learning to sustain dense prediction tasks, i.e., high-quality segmentations.

On completion of training, TOWER opts for the encoder $\boldsymbol{f}_\theta(\cdot)$ to initialize classification tasks. The encoder $\boldsymbol{f}_\theta(\cdot)$ and decoder $\boldsymbol{d}_\theta(\cdot)$ apply for initialization in the segmentation tasks.

In summary, the design enables correlated optimization of both the encoder and decoder, which are targeted on classification and segmentation in downstream tasks.

IV. Experiments and Results

Experiments were conducted (1) to evaluate TOWER's performance to recognize medical images in comparison with the state-of-the-art counterparts (Section IV-B), (2) to validate the effectiveness of TOWER via ablation studies based on medical image classification and segmentation (Section IV-C), and (3) to evaluate the capability of TOWER (Section IV-D).

TABLE I
AN OVERVIEW OF THE DATASETS USED IN THE EXPERIMENTS.

MedMNIST Dataset	Modality	Task	Dataset scale		
			Train	Valid	Test
PathMNIST	Colon Pathology	Multi-class (9) classify	89996	10004	7180
BreastMNIST	Breast Ultrasound	Binary-class (2) classify	546	78	156
DermaMNIST	Dermatoscope	Multi-class (7) classify	7007	1003	2005
RetinaMNIST	Fundus Camera	Multi-class (5) classify	1080	120	400
OrganMNIST	Abdominal CT	Multi-class (11) classify	34581	6491	17779

Abbr.†	Dataset	Task	Modality	Input size
VFS	DRIVE	Blood Vessels Segmentation	Fundoscopic	512×512
LXS	Montgomery	Lung Segmentation	X-ray	224×224
TXC	Shenzhen	Binary-class (2) Tuberculosis Classification	X-ray	224×224
DXC	CheXpert	Multi-label (5) Binary-class (2) Five thorax diseases Classification	X-ray	224×224
LCS	LiTS	Liver Segmentation	CT	512×512

†Abbreviation: the first letter denoted the object of interest, i.e., V denoted vessel, D denoted thorax diseases, L denoted live, and T denoted tuberculosis. The second letter denoted the modality, i.e., X denoted X-ray, F denoted Fundoscopic, and C denoted CT. The last letter denoted the task, i.e., C denoted classification, and S denoted segmentation.

A. Datasets and Experiment settings

This study conducted comparison experiments with the state-of-the-art methods on a self-supervised benchmark [20], including blood vessels segmentation from fundoscopic images (DRIVE [21], [22] [3]), lung segmentation from X-ray images (Montgomery [23][4]), liver segmentation from CT images (LiTS [24][5]), tuberculosis classification from X-ray images (Shenzhen [23][6]), thorax diseases classification from X-rays images (CheXpert [25][7])). Ablation study used the MedMNIST decathlon [26], [27] and DRIVE to validate the effectiveness of TOWER (see Table I). MedMNIST was a lightweight AutoML benchmark for medical image classification [8] and covered diverse data modalities, dataset scales, and tasks. This study selected the PathMNIST, BreastMNIST, DermaMNIST, RetinaMNIST, and OrganMNIS from MedMNIST.

Data augmentations of X-ray datasets (LXS, TXC, DXC) included random resizing cropping, random horizontal flipping, and random rotation. Data augmentations of fundoscopic dataset (VFS) included random rotation, gaussian noise, color dithering, and horizontal, vertical, and diagonal flipping. Data augmentations of CT dataset (LCS) included random rotation, gaussian noise, color dithering, and horizontal, vertical, and diagonal flipping. All the above datasets were split into the official training set and test set. All images were normalized.

For the classification tasks, TOWER was pre-trained with cross-entropy loss. The batch size was 128. The initial learning rate of pre-training was $1e^{-3}$ and it decreased to $2e^{-4}$ for fine-tuning. The fine-tuning process ended after 100 epochs or early stopped with patience 30. For the segmentation tasks, the loss function was standard pixel-wise cross-entropy loss. The batch size was 32. The initial learning rate of pre-training was $1e^{-2}$ and it decreased to $1e^{-3}$ for fine-tuning. The fine-tuning process ended after 200 epochs or early stopped with patience 30.

The run-time infrastructure for the experiments was mainly formed by PyTorch 1.10.0 with CUDA 10.2 over Four NVIDIA 1080Ti GPUs. The Adam optimizer [28] and cosine

[3]http://www.isi.uu.nl/Research/Databases/DRIVE
[4]www.kaggle.com/datasets/kmader/pulmonary-chest-xray-abnormalities
[5]https://competitions.codalab.org/competitions/17094
[6]http://openi.nlm.nih.gov/imgs/collections/ChinaSet_AllFiles.zip
[7]https://stanfordmlgroup.github.io/competitions/chexpert/
[8]https://medmnist.com/

TABLE II
TOWER OUTPERFORMED STATE-OF-THE-ART SELF-SUPERVISED PRE-TRAINING METHODS AND IMAGENET INITIALIZATION ON DIVERSE DOWNSTREAM TASKS. RESULTS STYLE: **BEST**, METHODS THAT OUTPERFORMED IMAGENET-BASED INITIALIZATION.

Pre-training methods	Classification tasks (AUC)		Segmentation tasks (Dice)		
	TXC	DXC	LCS‡	LXS	VFS
Random init.	89.03±1.82	86.62±0.46	92.75±0.57	97.55±0.36	78.27±0.40
ImageNet init.	95.62±0.63	87.10±0.36	94.19±0.18	98.19±0.13	79.20±0.34
InsDis	94.81±0.73	87.21±0.36	94.37±0.13	98.25±0.03	79.03±0.34
Model Genesis	95.91±0.63	87.79±0.47***	94.24±0.22	98.43±0.12	79.22±0.30
CMC	94.93±1.18	87.46±0.46	94.27±0.26	98.49±0.11***	79.50±0.45
MoCo-v1	94.54±0.42	86.98±0.11	93.87±0.72	98.08±0.14	78.98±0.45
PIRL	93.34±2.72	86.79±0.35	94.05±0.21	98.02±0.11	79.24±0.42
SimCLR-v1	94.45±0.76	87.66±0.14	93.54±0.18	98.19±0.10	79.00±0.18
MoCo-v2	95.57±0.90	86.94±0.20	93.88±0.21	97.79±0.50	79.23±0.19
SimCLR-v2	95.29±0.93	86.86±0.37	93.85±0.17	98.16±0.20	78.72±0.37
SeLa-v2	96.23±0.81	87.24±0.29	93.98±0.21	98.28±0.04	79.65±0.19
InfoMin	95.02±1.40	86.67±0.10	94.26±0.17	97.94±0.16	79.63±0.30
BYOL	94.69±0.78	87.09±0.40	93.76±0.25	98.20±0.08	79.39±0.22
DeepCluster-v2	96.09±0.68	87.01±0.19	93.84±0.25	98.24±0.05	79.66±0.21**
SwAV	95.72±0.50	87.06±0.50	93.63±0.31	98.28±0.05	79.65±0.14
PCL-v1	95.15±0.53	86.90±0.25	94.55±0.09*	98.25±0.05	78.99±0.21
PCL-v2	95.45±0.62	87.27±0.19	93.95±0.20	98.26±0.06	79.06±0.19
Barlow Twins	94.50±0.88	87.25±0.27	93.86±0.07	98.23±0.05	79.48±0.16
TOWER	**96.53±0.50**	**88.89±0.33**	**94.66±0.05**	**98.95±0.08**	**80.38±0.41**

‡ Reproduced by ourselves under the same protocol.
⋆ Statistical analysis between the **SECOND-BEST** method and our proposed **TOWER** was also conducted in each target task, where * denoted TOWER significantly outperformed the second-best method with p-value<0.005. ** denoted p-value<0.001. *** denoted p-value<0.0001.

learning rate decay schedulers were applied. The performance was measured in terms of AUC (%) and Dice (%) for classification and segmentation, respectively. All results were evaluated without post-processing and reported in (mean±std.) across ten independent trials.

B. Comparison with the state-of-the-art methods on a self-supervised benchmark

This section compared TOWER with InsDis [29], Model Genesis [13], CMC [30], MoCo v1-v2 [6], [7], SimCLR v1-v2 [5], [31], PIRL [32], PCL v1-v2 [33], SeLa v2 [8], [34], InfoMin [35], BYOL [36], DeepCluster v2 [8], [9], SwAV [8], Barlow Twins [37], ImageNet-based supervised initialization, and random initialization. The model was initialized by these pre-training methods and then fine-tuned on TXC, DXC, LCS, LXS, and VFS downstream tasks. This study also conducted statistical analysis between TOWER and the second-best self-supervised method in each target task.

Table II reported that TOWER could achieve significantly higher accuracy in all segmentation tasks, demonstrating the robustness of TOWER for promising initialization in dense prediction tasks. Specifically, for VFS, LXS, and LCS segmentation tasks, the backbone model initialized by TOWER achieved 0.73, 0.41, and 0.59 points increase in the Dice score, respectively, compared with the second-best method. For TXC and DXC classification tasks, the AUC scores of the encoder initialized by TOWER increased by 0.29 and 0.14 points over the second-best method, respectively.

Table II also indicated that the performances of most classic contrastive learning methods were inferior to the performances of ImageNet-based supervised pre-training method for medical image recognition, where MoCo-v1 did not perform as well as ImageNet-based initialization in all tasks, SimCLR-v1 underperformed the ImageNet-based initialization in TXC, LXS, and LCS.

C. Ablation study

Comprehensive ablation studies had been conducted to evaluate the design of individual components, basically via

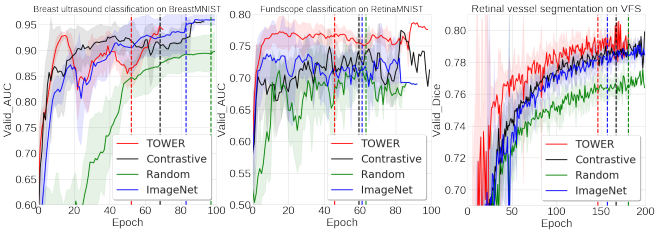


Fig. 6. The convergence curve during validation. The vertical red, blue, black, and green dash lines indicated the early-stopped average epochs of TOWER, ImageNet, SimCLR v1, and random initialized.

multi-modality medical image classification (MedMNIST) and retinal vessel segmentation (VFS).

Table III reported the performance between the individual components and the combined scheme. Fig. 6 showed the convergence curve during validation.

Additional experiments compared (1) the contrastive learning with reconstructive proxy tasks (denoted as $\mathcal{L}_{Con}^{NL+M}$) and (2) MoCo with classic data augmentations (denoted as $\mathcal{L}_{Con}^{moco}$). Experimental results indicated that:

(1) When applying the classic contrastive learning method (MoCo) to recognize medical images, we observed that the $\mathcal{L}_{Con}^{moco}$ did not perform as well as pre-trained ones from ImageNet in most studies. In contrast, the proposed contrastive learning component outperformed the classic one, i.e., the $\mathcal{L}_{Con}^{NL+M}$ outperformed the $\mathcal{L}_{Con}^{moco}$, demonstrating the effectiveness of reconstructive proxy tasks in enhancing contrastive learning.

(2) The results of the combined training scheme outperformed any single reconstructive proxy task, demonstrating the scalability and effectiveness of the proposed reconstructive proxy tasks in enhancing generative learning.

(3) The whole framework achieved the best performance and the fastest convergence of all components after synergizing contrastive and generative learning, demonstrating the effectiveness of complementary training of generative and contrastive learning.

D. Discussions

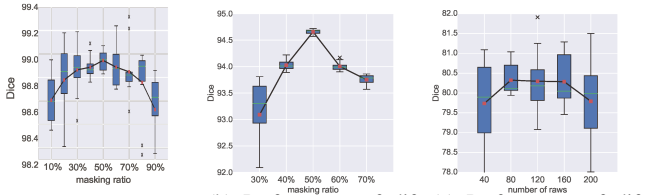
TOWER achieved the best performance in the benchmark (see Table II), which demonstrated that TOWER successfully enhanced contrastive learning by enriching the insufficient diversity (Issue #1), especially when medical images were annotation-free, without semantics, and without guarantee of quality. Table II also demonstrated that TOWER unleashed the power of 2D pre-training for medical image recognition, excelled in obtaining the significant representations demanded by a high-quality segmentation (Issue #2), and could be a preferred alternative for the ImageNet-based pre-training. This study attributed this performance superiority to the synergistic effect of contrastive and generative learning with a priori knowledge.

In addition, four metrics were highlighted to evaluate TOWER's capability, including: (1) the optimal masking ratios in masked reconstruction proxy tasks, (2) the distribution of extracted representations on medical image recognition, (3) the semantic-consistency between the encoder and decoder

TABLE III
PERFORMANCE OF THE PROPOSED SELF-SUPERVISED PRE-TRAINING COMPONENTS ON VARIOUS DOWNSTREAM TASKS. ↑: THE LARGER, THE BETTER. ↓: THE SMALLER, THE BETTER. RESULTS STYLE: **BEST**

Methods	$\mathcal{L}_{Con}^{moco}$	$\mathcal{L}_{Gen}^{NL}$	$\mathcal{L}_{Gen}^{M}$	$\mathcal{L}_{Con}^{NL+M}$	Classification PathMNIST	BreastMNIST	DermaMNIST	RetinaMNIST	OrganMNIST	Segmentation VFS
					AUC (%)					Dice (%)
Random					94.61±0.40	83.42±2.77	87.23±1.62	69.46±1.90	98.92±0.17	78.27±0.40
ImageNet					97.85±0.45	86.48±2.73	90.55±0.62	71.96±1.07	99.52±0.11	79.20±0.34
Contrastive	✓				96.93±0.66	87.80±2.49	90.04±0.42	69.98±1.89	99.51±0.06	79.23±0.19
TOWER		✓			97.46±0.57	86.23±3.65	90.49±0.93	71.52±1.22	99.49±0.06	79.82±0.39
			✓		97.79±0.54	85.58±2.99	90.13±0.68	71.20±1.34	99.49±0.07	79.60±0.54
		✓	✓		97.83±0.38	87.25±1.77	90.90±0.47	71.91±1.30	99.51±0.06	79.94±1.16
				✓	98.03±0.27	88.21±2.90	91.72±0.95	72.61±0.58	99.57±0.06	80.31±0.45
		✓	✓	✓	**98.49±0.24**	**88.40±1.04**	**91.78±0.30**	**73.01±0.94**	**99.60±0.04**	**80.38±0.30**

Note: $\mathcal{L}_{Con}^{moco}$ denoted the classic contrastive learning method (MoCo) without the proposed proxy task. $\mathcal{L}_{Gen}^{NL}$ denoted the generative learning method with nonlinear translation. $\mathcal{L}_{Gen}^{M}$ denoted the generative learning method with random masked reconstruction. $\mathcal{L}_{Con}^{NL+M}$ denoted the contrastive learning method with the proposed knowledge-based reconstructive proxy tasks. The results represented that $\mathcal{L}_{Con}^{NL+M}$ was statistically significantly better than the $\mathcal{L}_{Con}^{moco}$ with p-value <0.05. The results denoted TOWER was statistically significantly b



(a) Performance of different ratios of strip-wise masks on X-ray images. (b) Performance of different ratios of block-wise masks on CT images. (c) Performance of different ray number of rays-wise masks on fundoscopic images.

Fig. 7. An experiment of the impact of different masking ratios on downstream tasks. The boxplots showed fine-tuning results for recognizing medical images with different modalities and at different masking ratios. The red dot indicated the mean value across ten independent trials.

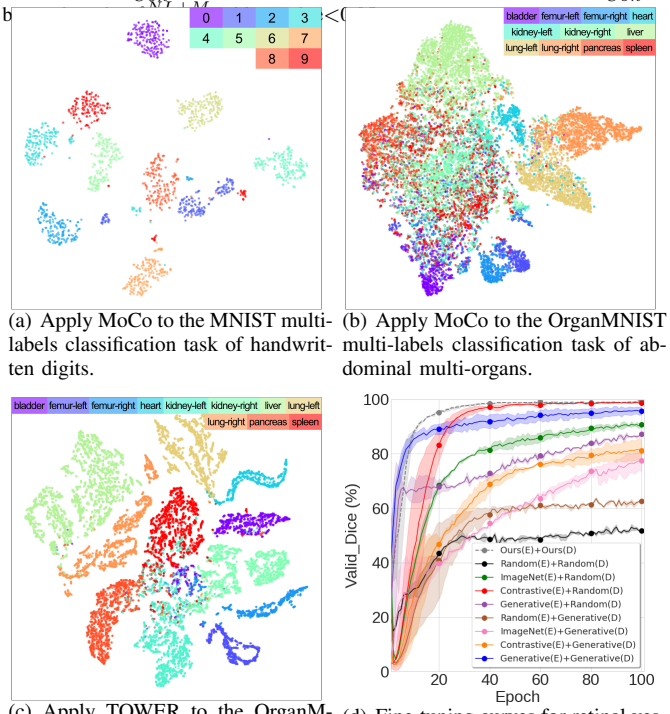
initialized by different pre-training methods, and (4) the label efficiency via semi-supervised experiments under different label fractions.

1) The optimal masking ratios for reconstruction: Fig.7 showed the performance of different masking ratios on three modalities of medical images. The optimal ratio varied for different kinds of masks. In this study, the best number of rays in the rays-wise mask was 80 for retinal images. In the block-wise mask for CT images and stripe-wise mask for X-ray images, the best masking ratios were 50%.

This finding was consistent with the understanding of the mask reconstruction proxy task. A low mask ratio led to a simple proxy task, and a high mask ratio led to insufficient image information. These improper mask ratios resulted in difficulties for the model to learn significant representations.

2) Visualization of the extracted robust representations: In order to explore the impact of the proposed restorative proxy tasks on contrastive learning, the representations extracted by the encoder were visualized in Fig.8(a)(b)(c) via t-SNE [38]. The classic contrastive learning method (MoCo) was selected as the baseline. This study conducted handwritten digits classification on MNIST [39] and multi-organs classification on OrganMNIST to evaluate the MoCo and TOWER.

MoCo was able to cluster representations of natural images correctly but failed in the medical image classification task. In contrast, the proposed knowledge-based restorative proxy tasks enriched the sample diversity of medical images (Issue #1), thus successfully constructing robust positive/negative pairs and correctly clustering the unlabelled representations.



(a) Apply MoCo to the MNIST multi-labels classification task of handwritten digits. (b) Apply MoCo to the OrganMNIST multi-labels classification task of abdominal multi-organs. (c) Apply TOWER to the OrganMNIST multi-labels classification task of abdominal multi-organs. (d) Fine-tuning curves for retinal vessel segmentation on LXS.

Fig. 8. Illustration of the extracted representations on different tasks (Zoom in for more details). Fig.(a) was trained on MNIST. Fig.(b) and Fig.(c) were trained on OrganMNIST. Fig.(d) reported the fine-tuning curves in segmentation with different pre-training methods, where (E) denoted a pre-trained encoder while (D) denoted a pre-trained decoder.

3) Semantic-consistent learning with different encoder-decoder-oriented pre-training schemes: In order to investigate the semantic-consistency of different encoder-decoder-oriented pre-training schemes on downstream tasks, this section conducted fine-tuning experiments with the following initialization schemes on VFS task: (1) Random initialization for encoder and decoder, (2) ImageNet-based pre-training of encoder [9], (3) Generative learning-based pre-training of encoder and decoder [10], (4) Contrastive learning-based pre-training of encoder [11], and (5) TOWER-based pre-training of

[9]Checkpoint was provided by `torchvision.models.resnet50(pretrained=True)`.

[10]Checkpoint was provided by Genesis [13]. Generative learning pre-trained a U-Net as encoder-decoder via loss function $\mathcal{L}_{Gen}^{NL}$.

[11]Checkpoint was provided by MoCo_CXR [40]. Contrastive learning pre-trained a ResNet50 as encoder via loss function $\mathcal{L}_{Con}^{moco}$.

encoder and decoder [12].

Fig.8(d) reported the fine-tuning curves with the encoder and decoder initialized by the above schemes. The proposed correlated optimization between the encoder and decoder achieved the best performance and the fastest convergence among all listed methods in Fig.8(d), demonstrating that TOWER unlocked the power of generative and contrastive learning and provided a significant representation for high-quality segmentation (Issue #2).

In contrast, the fine-tuning process of the generative learning-based pre-training method ([Generative(E)+ Generative(D)]) was sub-optimal in segmentation. Besides, the convergence of the model initialized by the classic contrastive learning method was hindered by the random initialization of the decoder, where the convergence speed of [Contrastive(E)+Random(D)] was slower than [TOWER(E)+TOWER(D)], demonstrating the significance of initializing the decoder for contrastive learning in dense pixel prediction tasks.

Note that when the encoder and decoder were initialized by different pre-training methods, it did not lead to improvements, but rather hindered performance and convergence, e.g., [Contrastive(E)+TOWER(D)] required more training epochs to obtain comparable performance to [Contrastive(E)+Random(D)]. The same was true for [ImageNet(E)+TOWER(D)]. In contrast, TOWER overcame this pitfall (Issue #2) and provided semantic-consistent initialization with better performance and faster convergence.

4) Label-efficient learning with few labelled samples: In order to explore the label-efficiency of TOWER on different percentages of labelled data, this section conducted semi-supervised experiments on PathMNIST, BreastMNIST, DermaMNIST, RetinaMNIST, and OrganMNIST. Fig.9 displayed the test AUC (%) of the model initialized from random, ImageNet, and TOWER under different label fractions. TOWER could mitigate the lack of annotations, resulting in label-efficient representation learning for medical image recognition. With decreasing amounts of labelled data, TOWER retained a much higher performance on all downstream tasks, especially when few labelled samples were available. Besides, TOWER could be fine-tuned on a few labelled datasets to achieve comparable performance to the full-labelled dataset. Specifically, compared to training from scratch, initialization with TOWER could reduce the annotation cost by 99%, 54%, 71%, 70%, and 92% for PathMNIST, BreastMNIST, DermaMNIST, RetinaMNIST, and OrganMNIST, respectively. Compared to ImageNet-based supervised pre-training, initializing with TOWER could reduce the annotation cost by 94%, 15%, 31%, and 20% for PathMNIST, BreastMNIST, DermaMNIST, RetinaMNIST, respectively. Compared to SimCLR v1-based self-supervised pre-training, initializing with TOWER could reduce the annotation cost by 96%, 52%, and 70% for PathMNIST, DermaMNIST, RetinaMNIST, respectively.

Overall, Tower[13] had made significant progresses towards solving the open issues (Section I): (1) Robust representations could be learned in an unsupervised manner from unlabelled medical images by enriching low diversity with a priori knowledge-based proxy tasks; (2) High-quality segmentation had been enabled by bridging the generative and contrastive learning to co-optimize the encoder and decoder.

[12]2D U-Net was pre-trained via loss function $\mathcal{L}_{Con}^{NL+M} + \mathcal{L}_{Gen}^{NL+M}$.

[13]The source code and the pre-trained models are available at https://github.com/lichen14/TOWER.

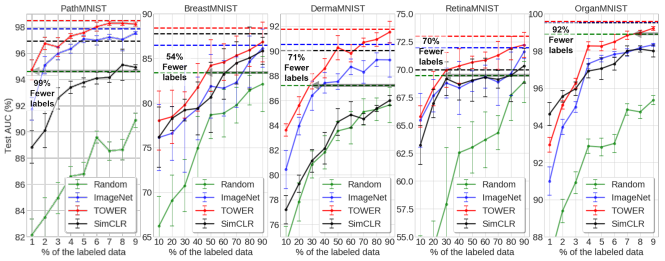


Fig. 9. Results of the semi-supervised experiments under label fractions. The horizontal red, blue, black, and green dash lines indicated the AUC on 100% labelled dataset that TOWER, ImageNet, SimCLR v1, and random initialized, respectively.

V. Conclusions

Aiming at the grand challenges for automated recognition of medical images towards clinical practices, this study developed a knowledge-based learning framework (TOWER). The framework synergizes generative learning and contrastive learning to enhance self-supervised learning towards high-quality initialization of AE models for reliable classification and segmentation tasks.

TOWER enabled sample space diversification via reconstructive proxy tasks to perform the nonlinear translation and random masked reconstruction based on a priori knowledge from clinic practices. The design supported enhanced representation learning of the representation of annotation-free images. Correlated optimization of the encoder and the decoder had been achieved by bridging contrastive and generative learning to serve the need for semantic segmentation.

Experimental results indicated that: (1) the proposed restorative proxy tasks could enrich the diversity of medical images and enhance contrastive learning with extended sample space; (2) TOWER could bridge generative and contrastive learning as a whole and provide better performance and faster convergence for segmentation by correlated-optimization between the encoder and decoder; (3) TOWER could mitigate the lack of annotations, resulting in label-efficient representation learning for medical image recognition (reduce up to 99% annotations in pathological classification).

Overall, TOWER significantly outperformed even the latest high-performance counterparts in terms of recognizing unlabelled medical images. It suggested that the key to sustaining reliable recognition of unlabelled medical images lied with appropriate portraying a priori knowledge in model optimization.

References

[1] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 43, no. 11, pp. 4037–4058, 2020.

[2] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk *et al.*, "Rethinking pre-training and self-training," *Advances in neural information processing systems (NeurIPS)*, vol. 33, pp. 3833–3845, 2020.

[3] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, early access, 2021.

[4] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations (ICLR)*, 2019.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning (ICML)*, 2020, pp. 1597–1607.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9729–9738.

[7] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[8] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9912–9924.

[9] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 132–149.

[10] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting self-supervised learning via knowledge transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9359–9367.

[11] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 577–593.

[12] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara *et al.*, "Self-supervised learning for medical image analysis using image context restoration," *Medical Image Analysis (MIA)*, vol. 58, p. 101539, 2019.

[13] Z. Zhou, V. Sodha, M. M. Rahman Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models genesis: Generic autodidactic models for 3d medical image analysis," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2019, pp. 384–393.

[14] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models genesis," *Medical Image Analysis (MIA)*, vol. 67, p. 101840, 2021.

[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollár *et al.*, "Masked autoencoders are scalable vision learners," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 000–16 009.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention (MICCAI)*, 2015, pp. 234–241.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2019.

[19] D. Levin, "Development of non-linear transformations for improving convergence of sequences," *International Journal of Computer Mathematics*, vol. 3, no. 1-4, pp. 371–388, 1972.

[20] M. R. Hosseinzadeh Taher, F. Haghighi, R. Feng, M. B. Gotway, and J. Liang, "A systematic benchmarking analysis of transfer learning for medical image analysis," in *MICCAI Workshop Domain Adaptation and Representation Transfer*, 2021, pp. 3–13.

[21] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever *et al.*, "Ridge-based vessel segmentation in color images of the retina," *IEEE Transactions on Medical Imaging (TMI)*, vol. 23, no. 4, pp. 501–509, 2004.

[22] M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, and M. D. Abramoff, "Comparative study of retinal vessel segmentation methods on a new publicly available database," in *Medical Imaging 2004: Image Processing*, vol. 5370, 2004, pp. 648 – 656.

[23] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh *et al.*, "Automatic tuberculosis screening using chest radiographs," *IEEE Transactions on Medical Imaging (TMI)*, vol. 33, no. 2, pp. 233–245, 2014.

[24] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis (MIA)*, p. 102680, 2022.

[25] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, vol. 33, no. 01, 2019, pp. 590–597.

[26] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," in *IEEE International Symposium on Biomedical Imaging*, 2021, pp. 191–195.

[27] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke *et al.*, "Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification," *arXiv preprint arXiv:2110.14795*, 2021.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference for Learning Representations (ICLR)*, 2015.

[29] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 3733–3742.

[30] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 776–794.

[31] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 22 243–22 255.

[32] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6707–6717.

[33] J. Li, P. Zhou, C. Xiong, and S. Hoi, "Prototypical contrastive learning of unsupervised representations," in *International Conference on Learning Representations (ICLR)*, 2021.

[34] A. YM., R. C., and V. A., "Self-labelling via simultaneous clustering and representation learning," in *International Conference on Learning Representations (ICLR)*, 2020.

[35] Y. Tian, C. Sun, B. Poole, D. Krishnan *et al.*, "What makes for good views for contrastive learning?" in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6827–6839.

[36] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch *et al.*, "Bootstrap your own latent - a new approach to self-supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 21 271–21 284.

[37] J. Zbontar, L. Jing, I. Misra, Y. Lecun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning (ICML)*, 2021, pp. 12 310–12 320.

[38] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[40] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco pretraining improves representation and transferability of chest x-ray models," in *Conference on Medical Imaging with Deep Learning (MIDL)*, vol. 143, 2021, pp. 728–744.