Pattern Recognition Letters 165 (2023) 154–160

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

GDRL: An interpretable framework for thoracic pathologic prediction

Yirui Wu[a,b,d], Hao Li[a,b], Xi Feng[c], Andrea Casanova[e], Andrea F. Abate[f], Shaohua Wan[g,*]

[a] Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing 210093, China
[b] College of Computer and Information, Hohai University, Nanjing 210093, China
[c] College of Harbor, Coastal and Offshore Engineering, Hohai University, Nanjing 210093, China
[d] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130015, China
[e] Department of Mathematics and Computer Science, University of Cagliari, Italy
[f] University of Salerno, Via Giovanni Paolo II, 132, Fisciano, Salerno 84084, Italy
[g] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, China

ARTICLE INFO

Article history:
Received 6 July 2022
Revised 4 December 2022
Accepted 23 December 2022
Available online 24 December 2022

Edited by Andrea F. Abate

MSC:
41A05
41A10
65D05
65D17

Keywords:
Disentangled representation learning
Group-disentangled feature representation
Thoracic pathologic prediction

ABSTRACT

Deep learning methods have shown significant performance in medical image analysis tasks. However, they generally act like “black box” without explanations in both feature extraction and decision processes, leading to lack of clinical insights and high risk assessments. To aid deep learning in envisioning diseases with visual clues, we propose a novel Group-Disentangled Representation Learning framework (GDRL). The key contribution is that GDRL completely disentangles latent space into disease concepts with abundant and non-overlapping feature related explanations, thus enhancing interpretability in feature extraction and decision processes. Furthermore, we introduce an implicit group-swap structure by emphasizing the linking relationship between semantical concepts of disease and low-level visual features, other than explicit explanations on general objects and their attributes. We demonstrate our framework on predicting four categories of diseases from chest X-ray images. The AUROC of GDRL on ChestX-ray14 for thoracic pathologic prediction are 0.8630, 0.8980, 0.9269 and 0.8653 respectively, and we showcase the potential of our framework in enhancing interpretability of the factors contributing to different diseases.

1. Introduction

Advances in deep learning has resulted in significant performance in medical image analysis tasks [1–3]. However, they generally act like “black box” without explanations in both feature extraction and decision processes, leading to lack of clinical insights and high risk assessments [4]. Most attempts to explain DL focus on ‘post-hoc’ analysis by proving the importance of low-level features (e.g., parts of an image) in producing accurate predictions [5]. However, they fail in linking low-level features with higher-level pathology concepts, and visually explain decision-making progress, where both are valuable for clinicians and patients to understand DL for pathology prediction.

As an alternative way, interpretable DL methods consider the need for knowledge related explanations into the structure designing. In other words, these methods are naturally transparent and interpretable by properly encoding knowledge in advance. In this way, in medical image analysis, we can locally tune the feature output of part of the model according to the clinical manifestations, instead of training the entire model from scratch, thus reducing the cost of model training. For instance, [6] propose an interpretable DL framework based on a variational auto-encoder (VAE) [7], which enables links between low-level features and higher-level explanatory concepts, as well as visualization of decision making boundary. Their work provides a significant step towards self-explaining DL methods. However, they achieve partly disentangled effects with overlapping and coarse-grained low-level features, resulting in confused explanations and low-accuracy classification results. The most important factor that distinguishes a partly disentangled latent space from a completely disentangled latent space is the overlap degree of the feature groups.

To completely disentangle the latent space, Group Supervised Learning (GSL) [8] designs an attribute swap operation to promote the consistency between latent representations and attributes, obtaining an excellent image synthesis effect by learning attributes from a group of samples. Following their idea, we propose a novel Group-Disentangled Representation Learning framework (GDRL) to aid deep learning in envisioning pathologys with visual clues. The

* Corresponding author.
E-mail addresses: wuyirui@hhu.edu.cn (Y. Wu), 211307030003@hhu.edu.cn (H. Li), xifeng@hhu.edu.cn (X. Feng), casanova@unica.it (A. Casanova), abate@unisa.it (A.F. Abate), shaohua.wan@uestc.edu.cn (S. Wan).

https://doi.org/10.1016/j.patrec.2022.12.020
0167-8655/

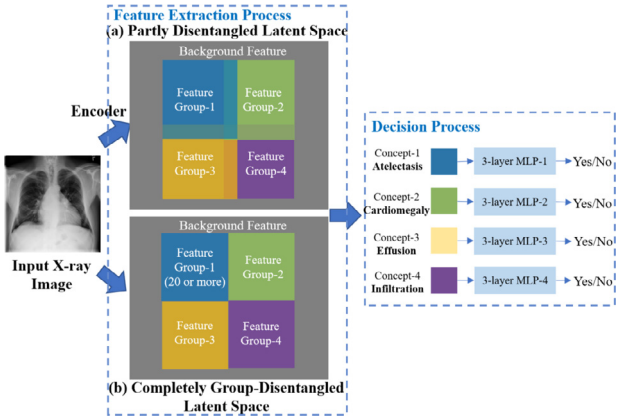

Fig. 1. Comparisons between partly and completely disentangled latent space, where the former one could be achieved by Esther et al. [6], and the latter one is generated by our GDRL.

key contribution is that GDRL completely disentangles latent space into pathology concepts with abundant and non-overlapping feature related explanations, thus enhancing interpretability in feature extraction and decision processes. Furthermore, we introduce an implicit group-swap structure by emphasizing the linking relationship between semantical concepts of pathology and low-level visual features, other than explicit explanations on general objects and their attributes.

Specifically, GDRL enables to extract representations of pathology concepts (e.g., Atelectasis, Cardiomegaly, Effusion, Infiltration, and Background in this paper) with a group of features, i.e., the number of features is up to 20 or more. To visually show how GDRL works, we draw in Fig. 1 to reveal the differences in latent representation space between GDRL and [6], thus offering clues on differences between completely and partly disentangled latent space, i.e., the overlaps of feature groups in partly disentangled latent space make the information of the corresponding pathology contained in feature groups not pure, resulting in the predictive performance of feature groups on the corresponding pathology being lower, while higher on other pathologies.

GDRL allows to decompose input (i.e., chest X-ray images) into a disentangled latent representation space with swappable components, each component encoding one pathology concept. With clinical knowledge on two samples sharing the identical latent values, i.e., the same pathology concept, an implicit group-swap structure is introduced, which seeks to link low-level visual features with high-level pathology concepts in space, thus laying an pathology interpretable basis in feature extraction process. In fact, the proposed implicit group-swap structure enforces semantic consistency of pathology concepts, and extracts features of pathology concepts by leveraging semantic links between samples, i.e., input chest X-ray images. Owning the ability to visualize pathology concepts with fine-grained and non-overlapping visual features, thus enhancing the interpretability of decision process of classifiers.

To sum-up, our contributions are as follows.

- We propose *Group-Disentangled Representation Learning* (GDRL) framework, which completely extracts group-disentangled pathology concept representations with fine-grained and non-overlapping features, thus promoting both interpretability and prediction accuracy.
- We introduce an implicit group-swap structure that enables to extract linking relationship between semantical concepts of pathology and low-level visual features in latent space.
- We experimentally demonstrate that GDRL can significantly improve classification accuracy compared with partly disentangled interpretable or other DL methods, and showcase the potential of GDRL to help clinicians' understanding in factors related with thoracic pathologic pathologys.

The rest of the paper is organized as follows. Section 2 gives an overview of the related work on relative aspects. In Section 3, details of the proposed GDRL, including Auto-Encoder network, group-Disentangled representation learning, disentanglement by Group-Swap module are discussed. Section 4 shows our experimental results, and finally Section 5 concludes the paper.

2. Related work

Disentangled Representation Learning. Deep learning has achieved significant improvements in multiple domains like digital twin [9], object detection [10], and so on. As one most promising aspect of deep learning, disentangled Representation Learning aims to learn disentangled representation for one specific task, where single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors.

Related methods can be divided into unit disentanglement methods and group-disentanglement methods. The former methods treat one feature as an independent concept. Following Variational Autoencoders, most unit-disentanglement methods achieve unit-disentanglement by incorporating KL-divergence into the objective to force the latent factors to be independent statistically [11].

Meanwhile, group-disentanglement methods treat a group of features as a concept. For example, [12] proposes a deep probabilistic model for learning a disentangled representation of a set of grouped observation, and separates the latent representation into semantically meaningful parts. Later, [13] proposes that image pairs within a group can generate partially swapped images by swapping their partial vectors.

Thoracic pathologic prediction. Early, Wang et al. [14] release a chest X-ray dataset, namely "ChestX-ray8", which are extracted from quantity of radiological reports, and benchmarked with different DLs pre-trained on ImageNet. Afterwards, CheXNet [4] has surpassed radiologists in its ability to predict pneumonia, which is proved by achieving high accuracy on "ChestX-ray8" dataset for pathologies prediction. Majoring in significant cybersecurity and privacy concerns, [15] focus their attention on the IEC 60 870-5-104 protocol, which is widely adopted in industrial healthcare systems, thus solving issues of threats on cyberattacks and intrusion detection with quantity of latest technologies like reinforcement learning and internet of medical things.

Later, CheXpert dataset [16], i.e., one of the largest CXR dataset, are proposed, which contains 200 studies and is manually annotated by 3 board-certified radiologists. Since rib segmentation based on chest X-ray images is essential in the computer-aided diagnosis systems of lung cancer, [17] propose a novel rib segmentation framework based on Unpaired Sample Augmentation and Multi-Scale Network, aiming to improve the accuracy of ribs segmentation with limited labeled samples. The classical random walk segmentation explores merely local affinity among neighboring pixels for cutting out objects, which falls short of effectiveness when handling distant repetitive patterns in medical domain. To alleviate the quandary, [18] propose to introduce nonlocal affinity among distant pixels with similar local features in the underlying segmentation graph, thus enabling label propagation among disconnected foregrounds and thus multiple repetitive patterns can be segmented jointly.

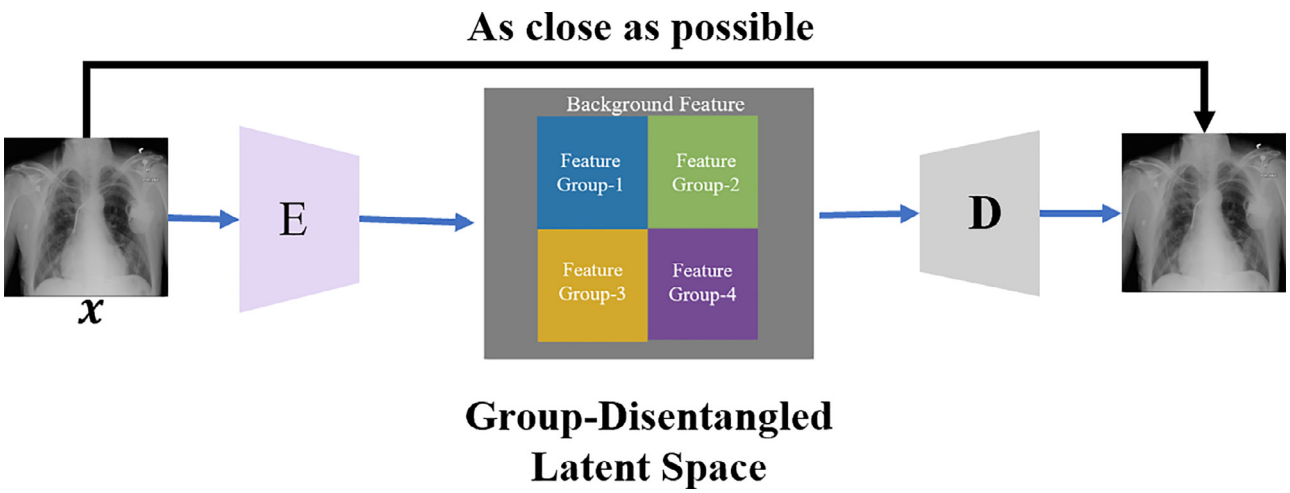

Fig. 2. Demonstration of encoding group-disentangled latent space based on auto-encoder. Each feature group in the latent space corresponds to a specific disease with medical usage.

3. Methodology

3.1. Auto-encoder network

Our GDRL framework is based on Auto-Encoder to decompose inputs (i.e., chest X-ray images) into a disentangled latent space. Therefore, before introducing our GDRL framework, we introduce the knowledge of auto-encoder network in this section.

The auto-encoder network is composed of an encoder network that embeds the input image into a latent space, followed by a decoder network trained to reconstruct the original image from the latent space. Typically we use convolutional and deconvolution neural networks as encoders and decoders respectively. We show these two processes in Fig. 2. After these processes is trained, we can take out the decoder and randomly pass in a code, hoping to generate a data similar to the original data through the decoder.

Formally, Auto-Encoder : $\mathcal{X} \to \mathcal{X}$ are composed of an encoder $E : \mathcal{X} \to \mathcal{R}^d$; and a decoder $D : \mathcal{R}^d \to \mathcal{X}$, where d denotes the dimension of the latent space $Z = E(X) \in \mathcal{R}^d$.

Although the auto-encoder network can compress the image into the latent space and reconstruct it well, we still do not know the internal structure of the latent vector, thus it lacks interpretability and can't create images arbitrarily by constructing latent vectors. Therefore, we propose GDRL based on auto-encoder network, which can completely disentangle the latent space into different parts, each part corresponding to a pathological concept or context.

3.2. Group-disentangled representation learning

To enhance interpretability in feature extraction, we wish to divide latent space into several semantic-specific parts. Formally, such property is defined as below:

Definition (Group-Disentangled Latent Space). *A group-disentangled latent space refers to a space consisting of several consecutive, non-overlapping subspaces, each of which is responsible for one specific concept.*

It can also be expressed in the view of row-vectors:

$$z^{(1)} = [g_1^{(1)}, g_2^{(1)}, \ldots, g_m^{(1)}, b^{(1)}], \tag{1}$$

where row-vector $z^{(1)}$ is the concatenation of m row-vectors $\{g_i^{(1)} \in \mathcal{R}^{d_i}\}_{i=1}^m$ and a background row-vector $b^{(1)} \in \mathcal{R}^b$ where $d = \sum_{i=1}^m d_i + b$, the values of $\{d_i\}_{i=1}^m$ and b are hyperparameters, and g_i corresponds to the concept c_i respectively.

To get group-disentangled latent space, we propose Group-Disentangled Representation Learning framework (GDRL), which extract group-disentangled representations of pathology concepts (e.g., Atelectasis, Cardiomegaly, Effusion, Infiltration, and Background) from a group of semantically related images and then uses them to accurately predict corresponding pathologys.

As shown in Fig. 3, the training step of our GDRL inputs a group of semantically-related images, then it trains the encoder and decoder through 3 modules, which enable the encoder to encode the image to a group-disentangled latent space.

The first module is a Linking Scheme, like auto-encoder, which links relationship between semantical concepts of pathology and low-level visual features, we calculate the reconstruction loss L_{ls} for each image. The next module is Group-Swap Module and will be introduced in Section 3.3, which enforces semantic consistency of pathology concepts by the after-swap reconstruction loss L_{gsm}. Group-Swap module and Group-cycle-swap module can retain the information of pathological concepts in the specified feature groups through swap operations, thus achieving completely group-disentangled latent space.

The last module is the prediction module (PM), which uses g_i to predict the value of concept c_i by constraint of binary cross entropy Loss L_{pm}. The prediction module consists of m 3-layer MLPs $\{M_i\}_{i=1}^m$, thus higher accuracy can reflect the informativeness of disentangled representations. We combine their losses into a total loss

$$\mathcal{L}(E, D, M; S) = L_{ls} + \lambda_{gsm} L_{gsm} + \lambda_{pm} L_{pm}, \tag{2}$$

where L_{gsm} and L_{pm} respectively are the losses of linking scheme, group-swap module and prediction module. Scalar coefficients $\lambda_{gsm}, \lambda_{pm} > 0$ control the relative importance of the loss term. And we can minimize the total loss $\mathcal{L}$ by gradient descent on parameters of encoder (E), decoder (D) and 3-layer MLPs.

The testing step of our GDRL uses the encoder to encode the input image into a group-disentangled latent space, and then predicts pathology concepts by prediction module.

For all experiments, the encoder E is composed of a convolutional layer, followed by 4 residual convolutional blocks with stride 2, followed by reshaping the response map to a vector, and finally a fully-connected layer to output 100-dim vector as latent feature. The decoder D mirrors the encoder, and is composed of a fully-connected layers, followed by reshape into cuboid, followed by 4 residual de-conv blocks with stride 2, then finally a de-conv layers to output a reconstruction image.

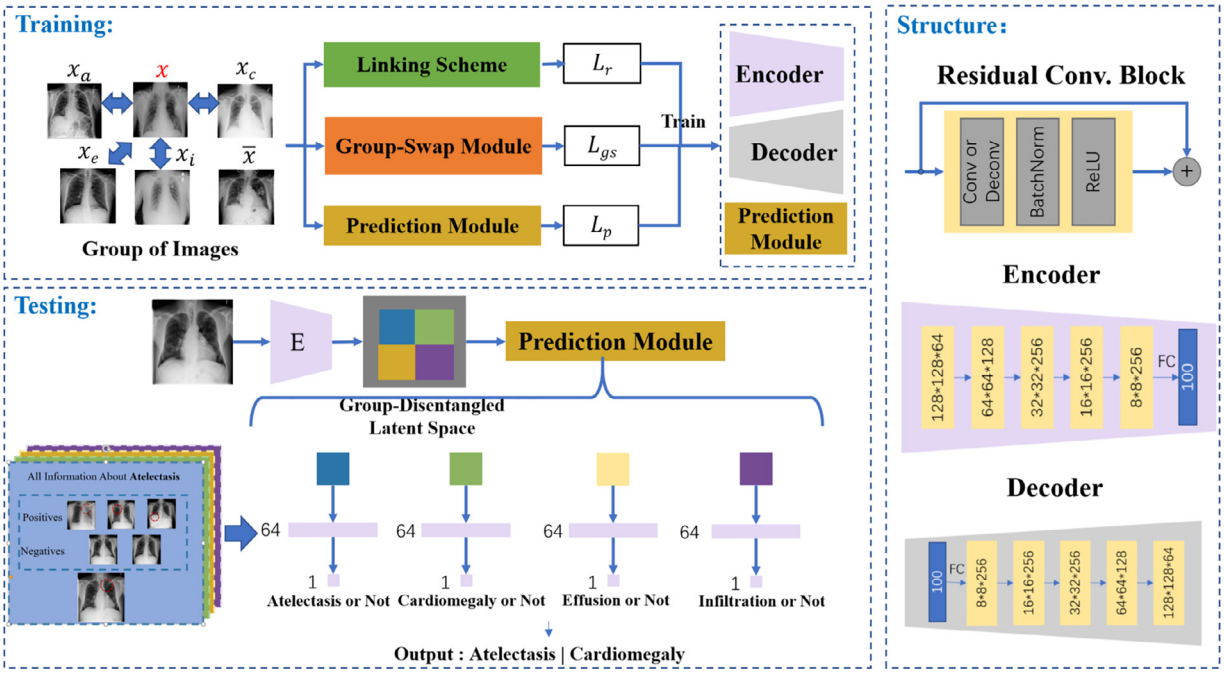

Fig. 3. Overall structure of GDRL, which trains an Auto-Encoder to extract group-disentangled representations of pathology concepts based on a group of semantically related images, and uses them to accurately predict corresponding concept labels.

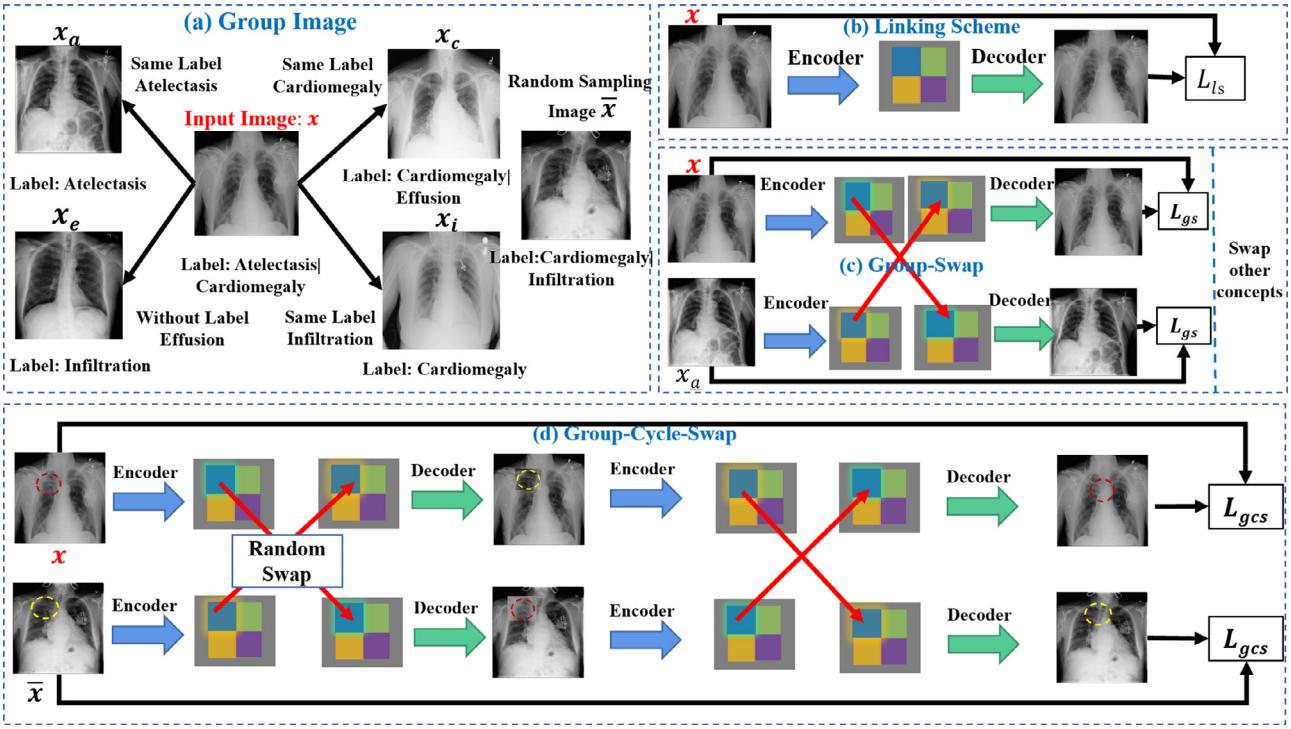

Fig. 4. Group-Swap Module for learning group-disentangled latent space. (a) **Group Image**, where we input a group of semantically related images to learn their common properties; (b) **Linking scheme**, where we calculate the self-reconstruction loss function for each image; (c) **Group-Swap**: For each pair of images, we swap part of the latent representations of their shared concept value. (d) **Group-Cycle-Swap**: For input image and a random sampling image, we encode, randomly swap subspace, then decode, re-encode, swap same subspace again for reversing the first swap, and decode to reconstruct the inputs.

3.3. Disentanglement by group-swap module

While training our GDRL, we wish to group-disentangle these latent spaces by E. Following the idea of Ge et al. [8], we propose an implicit group-swap structure to link low-level visual features with high-level pathology concepts in space. To achieve this, we design steps in Fig. 4. It's noted that the proposed structure is different from them in three aspects, where (1) we design a novel linking scheme to automatically search possible links between features and concept; (2) our group-swap structure prefer implicit and abstract feature aggregation for concept representation other than explicit attributes; (3) we consider other information (e.g. background) and disentangle it from concept representations, which may be entangled by them.

Group Image. As shown in Fig. 4(a), we first randomly select an image from the data set and randomly select images with the same labels as it, which implies that the latent subspace of these image pairs is similar. Then we randomly and singly selected an image for subsequent operations, and all these images add up to an image group.

Linking Scheme. As shown in Fig. 4(b), Linking Scheme is based on auto-encoder, which links relationship between semantical concepts of pathology and low-level visual features. Specifically, for each input X, we embed data in a low-dimensional vector by encoder. Then we link d_i units of the vector to a specific pathology concept c_i. Formally, we select a subset of the latent space $g_i = [\mu_{l_i+1}, \ldots \mu_{l_i+d_i}]$, where l_i is the start of the subset for concept c_i.

Group-Swap Module. We follow Ge et al. [8] and use two swap operations to enforces semantic consistency of pathology concepts, and extracts features of pathology concepts by leveraging semantic links between input images.

To simplify the notation to follow, **swap** operation is defined as:

$$\begin{aligned}&\textbf{swap}(z^{(1)}, z^{(2)}, k)\\ &= \textbf{swap}([\ldots, g_k^{(1)}, \ldots, b^{(1)}], [\ldots, g_k^{(2)}, \ldots, b^{(2)}], k)\\ &= [\ldots, g_k^{(2)}, \ldots, b^{(1)}], [\ldots, g_k^{(1)}, \ldots, b^{(2)}]\end{aligned} \tag{3}$$

As shown in Fig. 4(c), the first swap in our framework is called group-swap, considering an input image x and semantically relevant images in images group S, for all $x^o \in S,\ x^o \neq x$, the pair (x^o, x) share one concept value j (e.g., Atelectasis), we define a **Group-Swap** operation:

$$z, z^o = E(x),\ E(x^o) \text{ and } \mathrm{z_s}, \mathrm{z_s^o} = \textbf{swap}(\mathrm{z}, \mathrm{z^o}, \mathrm{j}) \tag{4}$$

and s.t.

$$L_{gs} = ||D(z_s) - x||_2^2 + ||D(z_s^o) - x^o||_2^2 \tag{5}$$

If sufficient sample pairs share only that concept for each concept, the loss of group-swap L_{gs} will be zero and group-disentanglement is achieved for that concept.

The second swap in our method is called group-cycle-swap, considering x and another randomly selected image $\bar{x}$, regardless of whether they share an concept value or not. As shown in Fig. 4(d), we first randomly choice a concept, then encode and swap the part of latent representation of the concept, after decode these swapped factor, we may have a partly changed images, which have no ground-truth images. After a same encode-swap-decode, outputs should be reconstructed into nearly the original images.

Formally, with the random chosen concept $j \sim U[1 \ldots m]$ the **Group-Cycle-Swap** operation can be defined as:

$$z, \bar{z} = E(x)\, E(\bar{x}),\ \text{ and } \mathrm{z_s}, \bar{\mathrm{z}}_\mathrm{s} = \textbf{swap}(\mathrm{z}, \bar{\mathrm{z}}, \mathrm{j}) \tag{6}$$

$$\hat{x} = D(z_s), \hat{\bar{x}} = D(\bar{z}_s) \tag{7}$$

$$\hat{z}, \hat{\bar{z}} = E(\hat{x})\, E(\hat{\bar{x}}),\ \text{ and } \hat{z}_\mathrm{s}, \hat{\bar{z}}_\mathrm{s} = \textbf{swap}(\hat{\mathrm{z}}, \hat{\bar{\mathrm{z}}}, \mathrm{j}) \tag{8}$$

and they are s.t.

$$L_{gcs} = ||D(\hat{z}_s) - x||_2^2 + ||D(\hat{\bar{z}}_s) - \hat{x}||_2^2 \tag{9}$$

Same as before, if sufficient samples are provided, the loss of group-cycle-swap L_{gcs} will be zero, then group-disentanglement is achieved. Finally, We combine the two group-swap losses into $L_{gsm} = L_{gs} + L_{gcs}$.

4. Experiments

We evaluate our method on its ability to learn group-disentangled representations and on its accuracy of thoracic pathologic prediction.

4.1. Datasets and measurements

We adopt two datasets to conduct thoracic pathologic prediction, i.e., chestxray-14 and ChestXpert, For former dataset, we select a subset for experiments, which contains 36,764 training images and 7353 testing images with 4 pathology labels (Atelectasis, Cardiomegaly, Effusion and Infiltration), which are extracting from the associated radiological reports using natural language processing. For the latter one, we also select a subset, which contains 162,188 training images and 32,437 testing images with 3 pathology labels(Pleural Effusion, Edema and Cardiomegaly).

To evaluate the performance of prediction, we follow the evaluation rules of both datasets, and adopt area under receiver operating characteristic curve (AUROC) as our evaluation metric.

4.2. Group-disentangled representation analysis

To see the effect of group-disentanglement of our GDRL, we use the subspaces of disease concepts to predict four thoracic pathologies by a simple 3-MLP. If the hidden subspace contains all the information about the disease, the predicted result should be a matrix with 1 on the diagonal and 0.5 on the rest.

We use Esther et al. [6] and standard auto-encoder with classification head as comparison methods. The former partly disentangles the latent space, and the latter is not a disentangled method. Table. 1 shows that GDRL successfully decomposes the image into a group-disentangled latent space and uses each subspace to accurately predict the corresponding concept, but not to predict other concepts. Results of two comparison methods, whose latent space is not completely group-disentangled, show that each subspace doesn't know what it corresponds to, so their AUROCs are nearly 0.5.

This result shows that the simple autoencoder structure cannot disentangle latent space. Instead, our method can effectively learn to group-disentangle representation and decompose the feature space into several independent parts, each of which represents a certain disease concept. However, other methods do not enforce the semantic consistency between the latent space and the concept of diseases, which leads to unsatisfactory results.

4.3. Accuracy of thoracic pathologic prediction

Table 2 shows that our GDRL has significantly improved on ChestXray-14 and ChestXpert dataset by prediction with group-disentangled latent representation compared with the existing methods and other disentangle methods. The blanks in Table 2 are due to the ChestXpert dataset has not yet been published in 2017, so it is hard to find the results of these methods on ChestXpert. For the ChestXray-14 dataset, the two methods compared were not tested on this dataset.

The proposed network obtains 86.30%, 89.80%, 92.69%, 86.53% for Atelectasis, Cardiomegaly, Effusion and Infiltration, being 5.36%, -2.68%, 6.31% and 13.08% higher than the second-highest achieved by CheXNet. Considering the reason for the decline of AUROC in predicting Cardiomegaly, we explored the ChestXray-14 and found that there was an extreme imbalance of the label of Cardiomegaly. This may be a weakness of interpretable models, making it difficult to learn concepts from these imbalanced datasets.

The accuracy of the proposed network is slightly lower on ChestXpert than the two latest networks, that is because our method considers not only the categories of predicted pathology, but also the interpretability of the network. It is useful in promoting clinicians' and patients' confidence and expanding usage of DL in automated pathology diagnosis.

Ablation Experiment To verify the effectiveness of each module in the proposed method, we conduct ablation studies on

Table 1
Group-Disentangled representation analysis. For each row of group-disentangled representations to predict each column of thoracic pathologies, we train a 3-layer MLP. The numbers in the table represent the AUROC of prediction on test datasets. Diagonals are bolded.

	GDRL				Puyol-Anton et al. [6] (partly disentangled)				AutoEncoder + PM (without disentangled)			
pathology concepts	Atel	Card	Effu	Infi	Atel	Card	Effu	Infi	Atel	Card	Effu	Infi
Atelectasis	**0.8630**	0.4855	0.5094	0.5005	**0.6136**	0.4960	0.4816	0.5050	**0.6076**	0.4990	0.4802	0.5297
Cardiomegaly	0.4822	**0.8980**	0.4836	0.5063	0.5062	**0.6610**	0.4968	0.4758	0.5067	**0.7048**	0.5183	0.4864
Effusion	0.4893	0.5061	**0.9269**	0.5229	0.5153	0.5038	**0.6688**	0.5099	0.4884	0.4985	**0.7444**	0.5292
Infiltration	0.4986	0.4900	0.4985	**0.8653**	0.4863	0.5230	0.5315	**0.5910**	0.4996	0.4955	0.4911	**0.6332**
Background	0.4983	0.5200	0.4926	0.4926	-	-	-	-	0.5045	0.5029	0.5087	0.4887

Table 2
Comparison Experiments. AUROC of Thoracic Pathologic Prediction by Different Methods on ChestXray-14 and ChestXpert. For each label approach, the highest AUROC scores are boldfaced. '-' means that the method is not evaluated on the data set.

	ChestXray-14				ChestXpert		
	Atel	Card	Effu	Infi	Effu	Edema	Card
GDRL	0.8630	0.8980	0.9269	0.8653	0.9	0.9023	0.8871
Rajpurkar et al. [4]	0.8094	**0.9248**	0.8638	0.7345	-	-	-
Yao et al. [19]	0.772	0.904	0.859	0.695	-	-	-
Wang et al. [14]	0.716	0.807	0.784	0.609	-	-	-
Ye et al. [20]	-	-	-	-	0.9166	0.9436	0.8703
Pham et al. [21]	-	-	-	-	**0.964**	**0.958**	**0.910**

Table 3
Ablation and Division Experiments. AUROC of Thoracic Pathologic Prediction by GDRL with different structure on ChestXray-14.

	Atel	Card	Effu	Infi
GDRL	**0.8630**	0.8980	**0.9269**	**0.8653**
$L_{ls} + L_{gs} + L_{pm}$	0.5780	0.6455	0.7122	0.6047
$L_{ls} + L_{pm}$	0.6076	0.7048	0.7444	0.6332
$L_{ls} + L_{gs} + L_{gcs}$	0.5065	0.4711	0.5032	0.5289
Less Background	0.8497	0.8749	0.9013	0.8633
More Background	0.8263	**0.9048**	0.8883	0.8445

ChestXray-14, where performance is listed in Table. As shown in Table 3, the AUROC will decrease by a large percentage without the help of L_{csr} module, since group-cycle-swap implies that swapping one attribute does not destroy latent information for other attributes. If we continue to remove the L_{sr} module, the model will degenerate to AutoEncoder+PM, the AUROC increases a little bit instead, it shows that AE+PM only need to focus on prediction task, but the entangled latent space still leads to poor prediction results.

If only L_p is removed, the AUROC of our model for the binary classification task is close to 50%, which is equivalent to not working. In a word, all these characteristics of the proposed method thus lead to being more accurate.

Parameter Setting Experiment As shown in the last two lines of Table. 3, considering that these thoracic pathologies are independent of each other, we distribute their corresponding latent subspace with same size. It's noted that less background of GDRL represents $g_i = 22, i = 1, 2, 3, 4$ and $b = 12$, and more background represents $g_i = 15, i = 1, 2, 3, 4$ and $b = 40$.

As we allocate less (as 12 in our paper) dimensions of latent space to represent background, the AUROC decreases by 1.33%, 2.31%, 2.56% and 0.2% for each pathology. If more (as 40 in our paper) latent space are used for background, the AUROC change by a percentage of -6.37%, +0.68%, -3.86% and -2.08%. Note that the change of Cardiomegaly is different from other pathologies. After exploring the data set, we found that this pathology had a serious imbalance in the data set, resulting in higher performance than usual. To summarize, this experiment shows that equally division is the most effective for this task.

4.4. Implement details

All our experiments were conducted on a server with two Intel Xeon E5-2620 v4 (@2.1GHz) CPUs and 4 NVIDIA GTX1080Ti graphic cards. Our experimental codes are mainly based on the PyTorch framework. Our initial learning rate is set as 0.0001, weight decay is 0.0001 and the momentum is 0.9. Due to the linear warm up mechanism, the learning rate increases from 1/30 to 0.01 in the first 500 iterations.

By default, both scalar coefficients λ_{gsm} and λ_{pm} are set to 1. All the proposed modules are added, and the latent space is equally divided, e.g., if $d = 100$, then $d_i = 20, i = 1, 2, 3, 4$, and $b = 20$. For comparison purposes, the start positions for concepts are evenly distributed, i.e., $l_1 = 0$, $l_2 = 20$, $l_3 = 40$, and $l_4 = 60$.

5. Conclusion

This paper proposes a Group-Disentangled Representation Learning (GDRL) framework, which completely extracts group-disentangled pathology concept representations with fine-grained and non-overlapping features, thus promoting both interpretability and prediction accuracy. GDRL makes the model truly understand the semantic information of the data. With clinical knowledge on two samples sharing the identical concept values, an implicit group-swap structure is introduced, which seeks to link low-level visual features with high-level pathology concepts in the latent space, thus laying an pathology interpretable basis in feature extraction process. According to the experimental results, the feature disentangling effect of the model can be improved by introducing more constraints to force the information to be retained in the corresponding feature groups. In addition, it is worth thinking about how to balance interpretability with the performance of downstream tasks.

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.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by a grant from National Key R&D Program of China under Grant No. 2021YFB3900601, the Fundamental Research Funds for the Central Universities under Grant B220202074, and Joint Foundation of the Ministry of Education (No. 8091B022123).

References

[1] Y. Han, T. Zhuo, P. Zhang, W. Huang, Y. Zha, Y. Zhang, M.S. Kankanhalli, One-shot video graph generation for explainable action reasoning, Neurocomputing 488 (2022) 212–225.
[2] A. Meiseles, D. Paley, M. Ziv, Y. Hadid, L. Rokach, T. Tadmor, Explainable machine learning for chronic lymphocytic leukemia treatment prediction using only inexpensive tests, Comput. Biol. Med. 145 (2022) 105490.
[3] O. Trigueros, A. Blanco, N. Lebeña, A. Casillas, A. Pérez, Explainable ICD multi-label classification of EHRs in spanish with convolutional attention, Int. J. Med. Inform. 157 (2022) 104615.
[4] P. Rajpurkar, J. Irvin, et al., CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning, CoRR (2017) abs/1711.05225.
[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: CVPR, 2016, pp. 2921–2929.
[6] E. Puyol-Antón, C. Chen, J.R. Clough, et al., Interpretable deep models for cardiac resynchronisation therapy response prediction, in: MICCAI, Vol. 12261, 2020, pp. 284–293.
[7] D.P. Kingma, M. Welling, Auto-encoding variational bayes, ICLR, 2014.
[8] Y. Ge, S. Abu-El-Haija, G. Xin, L. Itti, Zero-shot synthesis with group-supervised learning, ICLR, 2021.
[9] Y. Wu, H. Cao, G. Yang, T. Lu, S. Wan, Digital twin of intelligent small surface defect detection with cyber-manufacturing systems, ACM Trans. Internet Technol. (2022).
[10] Y. Wu, H. Guo, C. Chakraborty, M. Khosravi, S. Berretti, S. Wan, Edge computing driven low-light image dynamic enhancement for object detection, IEEE Trans. Netw. Sci. Eng. (2022), doi:10.1109/TNSE.2022.3151502. 1–1
[11] R. Gu, Y. Chen, S. Liu, H. Dai, G. Chen, K. Zhang, Y. Che, Y. Huang, Liquid: intelligent resource estimation and network-efficient scheduling for deep learning jobs on distributed gpu clusters, IEEE Trans. Parallel Distrib. Syst. 33 (11) (2021) 2808–2820.
[12] D. Bouchacourt, R. Tomioka, S. Nowozin, Multi-level variational autoencoder: learning disentangled representations from grouped observations, in: AAAI, 2018, pp. 2095–2102.
[13] A. Szabó, Q. Hu, T. Portenier, M. Zwicker, P. Favaro, Understanding degeneracies and ambiguities in attribute transfer, in: ECCV, Vol. 11209, 2018, pp. 721–736.
[14] X. Wang, Y. Peng, et al., ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: CVPR, 2017, pp. 3462–3471.
[15] P. Radoglou-Grammatikis, K. Rompolos, P. Sarigiannidis, V. Argyriou, T. Lagkas, A. Sarigiannidis, S. Goudos, S. Wan, Modeling, detecting, and mitigating threats against industrial healthcare systems: a combined software defined networking and reinforcement learning approach, IEEE Trans. Ind. Inf. 18 (3) (2021) 2041–2052.
[16] J. Irvin, P. Rajpurkar, et al., CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison, in: AAAI, 2019, pp. 590–597.
[17] H. Wang, D. Zhang, S. Ding, Z. Gao, J. Feng, S. Wan, Rib segmentation algorithm for x-ray image based on unpaired sample augmentation and multi-scale network, Neural Comput. Appl. (2021) 1–15.
[18] L. Wang, M. Li, X. Fang, M. Nappi, S. Wan, Improving random walker segmentation using a nonlocal bipartite graph, Biomed. Signal Process. Control 71 (Part) (2022) 103154.
[19] L. Yao, E. Poblenz, D. Dagunts, et al., Learning to diagnose from scratch by exploiting dependencies among labels, CoRR (2017) abs/1710.10501.
[20] W. Ye, J. Yao, H. Xue, Y. Li, Weakly supervised lesion localization with probabilistic-cam pooling, CoRR (2020) abs/2005.14480.
[21] H.H. Pham, T.T. Le, D.Q. Tran, D.T. Ngo, H.Q. Nguyen, Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels, Neurocomputing 437 (2021) 186–194.