Anatomy-Aware Self-Supervised Learning for Aligned Multi-Modal Medical Data

Hongyu Hu[1]
mathewcrespo@sjtu.edu.cn
Tiancheng Lin[1]
ltc19940819@sjtu.edu.cn
Yuanfan Guo[1]
gyfastas@sjtu.edu.cn
Chunxiao Li[2]
lcx198910@126.com
Rong Wu[2]
wurong7111@163.com
Yi Xu*[1]
xuyi@sjtu.edu.cn

[1] MoE Key Lab of Artificial Intelligence
Shanghai Jiao Tong University
Shanghai, China

[2] Department of Ultrasound
Shanghai General Hospital
Shanghai Jiao Tong University School of Medicine
Shanghai, China

Abstract

Consistency of anatomical structure naturally exists among medical images from multiple modalities, which provides powerful supervisory signals to self-supervised learning on aligned multi-modal medical images. However, it would lose efficacy due to modality-specific attributes when directly applying current pixel-wise or region-wise contrastive learning methods to pull aligned multi-modal data together in embedding space. To address this issue, we propose a novel anatomy-aware self-supervised learning framework, which represents anatomical structure in each modality using spatial similarity distribution between image patches, to alleviate the ill effects of modality-specific attributes and obtain a modality-consistent representation of anatomical structure. Significantly, we construct a correlation matrix to represent spatial similarity distribution and design a consistency loss to align the distributions across modalities to maintain anatomical consistency. Furthermore, we integrate it with instance-level discrimination into a unified contrastive framework, where the learned features are augmentation-invariant and modality-consistent. Extensive experiments on two medical datasets for the diagnosis of breast cancer and retinal diseases demonstrate that our proposed method achieves superior performance to current related work.

1 Introduction

Self-supervised learning (SSL) learning has emerged as an effective method for learning good feature representations. It leverages the input data itself as supervision [16, 22], such as context-instance relationships [7, 9, 17], instance-instance contrast [6, 11, 14, 24, 33], or dense contrast [23, 31, 34]. This characteristic is of great help to the field of medical

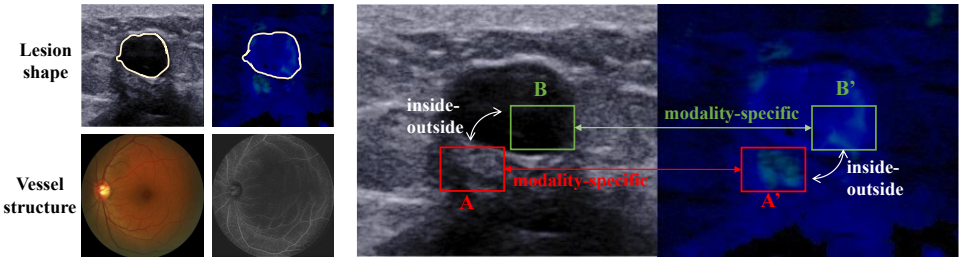
imaging, since labelling medical images is expensive, time-consuming and requires expertise [28]. Medical images are aligned or paired across multiple modalities, in which the same lesion or tissues are located in the same position under different imaging techniques, e.g., B-mode ultrasound and shear wave elastography (SWE), color fundus and fundus fluorescein angiography (FFA), etc. These modalities are complementary for a more accurate diagnosis, which has been proved in supervised learning paradigms [5, 10, 21, 36], but still understudied in self-supervised learning. Given that multi-modality medical images naturally provide more views than uni-modality data, how to effectively utilize such information as self-supervision is a key factor in self-supervised learning for multi-modal data.

Motivated by general paradigms in self-supervised learning for uni-modality data, many current works on multi-modality medical data mainly utilize general semantic correspondence as self-supervision, attracting different modalities of the same object in feature space. Holmberg *et al*. [15] suggested that practical pretext tasks in medical domain should be disease-related. Hence, they developed a novel pretext task, which employed two different modalities, including optical coherence tomography scans (OCT) and infrared fundus images, to predict retinal thickness. Li *et al*. [20] proposed to learn modality-invariant features and patient-similarity features in a contrastive learning for retinal disease diagnosis on paired FFA and color fundus images. The above-mentioned multi-modality methods all focus on overall semantic correspondence, ignoring local anatomical structures embedded in medical data. Since human organs or tissues are intrinsically structured, there is an inherent consistency underlying their appearance and layout in medical images [35](see Fig.1(a)).

With regard to local anatomy, an intuitive and direct solution is to transfer dense contrastive learning methods [23, 31, 32, 34] to multi-modality medical data. However, directly applying these methods would be sub-optimal, as they simply pull corresponding regions closer in feature space where modality-specific attributes would incur a strong bias in feature computation [12, 13]. For example, B-mode ultrasound reflects lesion shape [18], while SWE focuses more on tissue stiffness [19]. It is required to obtain a modality-consistent representation of anatomical structures with tolerance to modality-specific attributes.

To this end, we propose a novel anatomy-aware self-supervised learning framework for aligned multi-modality medical images in an integrated contrastive learning manner, to exploit the spatial similarity distribution across local patches as well as the commonly-used global information. For anatomical consistency, we construct a correlation matrix to represent spatial similarity distribution within each modality, and align the distribution across modalities to capture anatomical consistency. As is shown in Fig.1(b), patch A and B represent peritumoral and intratumoral areas respectively. It is known that biological changes in tumor-adjacent areas are potential predictive and prognostic markers to tumor diagnosis [27], which remains modality-consistent. Correspondingly, it is rational to compute similarity between two local patches and use the spatial similarity distribution to reflect the variations among the local anatomical structures. Significantly, such anatomical consistency is proposed to serve as a more reliable and robust cross-modality self-supervision. It is not only a soft regularization of local and anatomical correspondence to tolerate fine-grained modality-specific attributes, but also models the overall structure of patch interrelationships. For global representations, apart from augmented views of each modality, cross-modality views are also formulated as positive pairs, providing enhanced semantic diversities.

Extensive experiments are conducted on two aligned multi-modality medical datasets for the diagnosis of breast cancer and retinal diseases. Experimental results demonstrate that our proposed method achieves superior performance to current representative works in self-supervised learning, indicating that exploring spatial similarity distribution for modality-



(a) Anatomical similarity in different modalities (b) Relative region-wise relationship

Figure 1: Anatomical structure in each modality is represented using spatial similarity distribution between image patches, to alleviate the ill effects of modality-specific attributes and obtain a modality-consistent representation of anatomical structure.

consistent anatomical representation could further enhance the self-supervision signal. Ablation studies are carried out to further validate its effectiveness.

2 Methodology

2.1 Overall Framework

The proposed self-supervised learning framework is displayed in Fig.2, which explores the spatial similarity distribution as modality-consistent representation of anatomical structure. In the branch of self-supervision of anatomical consistency, we construct a correlation matrix to represent patch similarity distribution and design a consistency loss to align similarity distribution across modalities. In the branch of self-supervised representation learning of global-invariant features, modality-consistent and augmentation-invariant features are learned in a contrastive manner. The network is optimized by consistency loss and global contrastive loss simultaneously.

2.2 Problem Definition

Set N of paired data from two aligned modalities M_A, M_B, together with its augmented views $\hat{M}$ are fed into the neural network within a batch:

$$M = \{(m_A^1, m_B^1), (m_A^2, m_B^2), \ldots, (m_A^N, m_B^N)\}; \hat{M} = \{(\hat{m}_A^1, \hat{m}_B^1), (\hat{m}_A^2, \hat{m}_B^2), \ldots, (\hat{m}_A^N, \hat{m}_B^N)\} \quad (1)$$

The neural network G_θ consists of ℓ stacked convolutional layers θ as the backbone, followed by a projection head Θ:

$$G_\theta = G(M, \hat{M}; \theta_1, \theta_2, \ldots \theta_\ell, \Theta) \quad (2)$$

Our goal is to learn a good feature embedding network G_θ in an unsupervised manner, which can embed image m_A^i and m_B^i into highly distinguishable vectors f_A^i and $f_B^i \in \mathbb{R}^d$, where d denotes the embedding dimension.

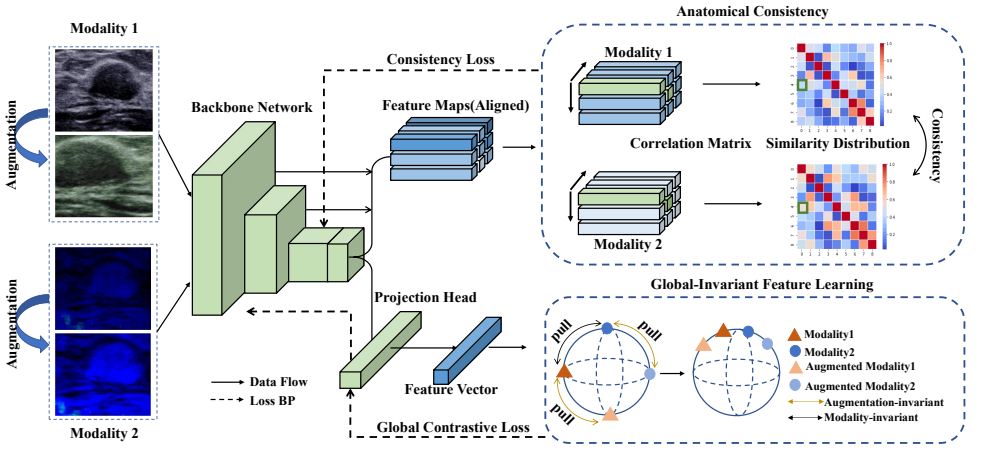


Figure 2: An overview of our proposed method. Feature maps of aligned multi-modality images are fed into the self-supervision of anatomical consistency. Feature vectors of aligned images, together with their augmentations go into the self-supervised learning of global-invariant features.

2.3 Self-Supervision of Anatomical Consistency

For the input pair data of multi-modality, modality-specific features are naturally embedded in different modalities, since they present different attributes [18, 19] and some tissues or tiny anatomical structures are only presented in a certain modality. Simply pulling region features from different modalities closer to pursue absolute anatomical consistency would incur strong bias. To alleviate the ill effects of these modality-specific attributes and motivated by the fact that relative region-wise relationship remains modality-consistent (see Fig.1(b) and Sec.1), we propose to model such relation using spatial similarity distribution to obtain modality-consistent representation of anatomical structures for cross-modality self-supervision. Moreover, it models overall structure of region interrelationships, while previous works [23, 31, 34] only focus on discriminating corresponding pixels or regions.

Feature Maps and Patches: When the input data M and $\hat{M}$ are fed into the neural network G_θ, for each convolutional layer θ, it produces a feature map of pixel-wise embedding v with the shape of $C \times H \times W$ for each single image (C denotes the number of channels, H and W are the height, width of the feature map). To get patch-wise (region-wise) features, we evenly divide the feature map into $N^h \times N^w$ patches with the shape of $\lfloor \frac{H}{N^h} \rfloor \times \lfloor \frac{W}{N^w} \rfloor$. The embedded feature s_i of the patch p_i is defined as:

$$s_i = \frac{\sum_{(a,b)\in p_i} v(a,b)}{||\sum_{(a,b)\in p_i} v(a,b)||_2} \tag{3}$$

where a, b denote the position coordinates of the feature map within the patch p_i.

Anatomical Consistency by Aligning Spatial Similarity Distribution: We first construct a correlation matrix A in Eq.4, to reflect spatial similarity distribution in the high-dimensional embedding space. Its elements denote patch-wise similarities.

$$A_{i,j} = sim(p_i, p_j) = s_i^T s_j, (i, j \in N^h \times N^w) \tag{4}$$

where $sim(\cdot)$ calculates the correlation score (cosine similarity) between two patches. In aligned pairs of multi-modality images, such similarity distribution should be pulled to be consistent across modalities. Therefore, we design a consistency loss to align this similarity distribution for anatomical consistency, which is defined as:

$$L_{SD} = \frac{1}{(N^h N^w)^2} \sum_j \sum_i (A_{i,j}^{(M_A)} - A_{i,j}^{(M_B)})^2 \tag{5}$$

where $A^{(M_A)}$ and $A^{(M_B)}$ denote the correlation matrix in corresponding modality respectively, $(N^h N^w)^2$ is the number of elements in the matrix. Note that in the network G_θ, hierarchical feature maps are produced by stacked layers of convolution θ. Therefore, the consistency of spatial similarity distribution widely exists in the backbone network, and the overall consistency loss is defined as:

$$L_C = \frac{1}{|Q|} \sum_{\theta_i \in Q} L_{SD}^{(\theta_i)} \tag{6}$$

where Q is a set of convolutional layers selected for the calculation of consistency loss ($|Q|$ denotes set size) and $L_{SD}^{(\theta_i)}$ is the consistency loss in the ith layer.

2.4 Global-Invariant Feature Representation

After the projection head Θ, the network G_θ embeds the input data (m_A, m_B) and $(\hat{m_A}, \hat{m_B})$ into high-dimensional feature vectors $(\mathbf{f_A}, \mathbf{f_B})$ and $(\hat{\mathbf{f}}_\mathbf{A}, \hat{\mathbf{f}}_\mathbf{B})$. We then normalize all the feature vectors by l_2 normalization, i.e., $||\mathbf{f_A}||_2 = ||\mathbf{f_B}||_2 = ||\hat{\mathbf{f}}_\mathbf{A}||_2 = ||\hat{\mathbf{f}}_\mathbf{B}||_2 = 1$. The overall global feature representations are then learned in a contrastive manner, which mines invariant representations across augmentations and modalities.

Augmentation-Invariant Features: The basic diagnosis of a medical image would not change under augmentations. Accordingly, feature representation should be robust enough to image augmentations. It can be implemented for each modality in a contrastive manner, where original images and their corresponding augmentation versions are positive pairs. The contrastive loss for augmentation-invariant features is defined as:

$$L_{AUG}^{M_A} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathbf{f_A^i} \cdot \hat{\mathbf{f}}_\mathbf{A}^\mathbf{i}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{f_A^i} \cdot \hat{\mathbf{f}}_\mathbf{A}^\mathbf{j}/\tau)}; \quad L_{AUG}^{M_B} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathbf{f_B^i} \cdot \hat{\mathbf{f}}_\mathbf{B}^\mathbf{i}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{f_B^i} \cdot \hat{\mathbf{f}}_\mathbf{B}^\mathbf{j}/\tau)} \tag{7}$$

$$L_{AUG} = L_{AUG}^{M_A} + L_{AUG}^{M_B} \tag{8}$$

where N denotes the size of the sample batch, dot product $\cdot$ is used to calculate cosine similarity and τ is the temperature parameter.

Modality-Invariant Features: Similar to augmentation invariant features, images of the different modalities from the same patient would share the same medical label and similar semantic information. Therefore, in a contrastive learning method, images of modality A and B belonging to the same patient are re-formulated as positive pairs, while the ones from different patients are considered to be negative samples. In practice, we treat each modality as an anchor and enumerate over the other, then add them up as a two-view modality-invariant loss:

$$L_M^{A \to B} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathbf{f_A^i} \cdot \mathbf{f_B^i}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{f_A^i} \cdot \mathbf{f_B^j}/\tau)}; \quad L_M^{B \to A} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathbf{f_B^i} \cdot \mathbf{f_A^i}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{f_B^i} \cdot \mathbf{f_A^j}/\tau)} \tag{9}$$

$$L_M = L_M^{A\to B} + L_M^{B\to A} \tag{10}$$

In general, the final global contrastive loss function to capture global invariant feature representations is defined as:

$$L_{Glo} = L_{AUG} + L_M \tag{11}$$

Our final learning objective is to optimize the overall loss function defined as:

$$L = L_C + \alpha L_{Glo} \tag{12}$$

where α is a scaling factor to balance different loss terms.

3 Experiments

Table 1: Experiment results of linear classification and finetuning on Breast US-SWE and Synthesized Retinal Fundus-FFA datasets. The best two metrics in each group of experiments are highlighted in red and blue. (Unit:%)

Pretrain Dataset	Evaluation Protocol	Method	Pretrain Dataset			Transfer Dataset		
			AUC	Acc	F1-score	AUC	Acc	F1-score
US-SWE	Linear Classification	InstDis-US	81.77	75.20	69.77	78.17	74.03	60.05
		InstDis-SWE	85.85	78.16	72.23	74.88	71.56	54.34
		CMC	86.77	79.89	74.57	79.12	74.17	58.61
		SimCLR	88.54	77.77	72.59	79.80	73.70	59.11
		InstDis-All	88.27	80.88	77.26	79.04	73.87	57.97
		DenseCL	87.62	80.95	76.89	79.56	75.27	62.81
		Ours	90.11	82.43	78.46	82.02	76.81	62.46
	Finetuning	w/o pretrain	83.15	74.86	74.97	85.11	79.29	66.97
		InstDis-US	84.87	78.07	73.04	87.13	81.75	69.24
		InstDis-SWE	89.32	82.21	78.42	86.99	81.75	70.81
		CMC	91.10	83.83	79.66	87.53	81.90	70.82
		SimCLR	90.22	81.76	79.34	88.94	80.06	68.16
		InstDis-All	88.27	81.23	77.76	88.40	81.60	70.39
		DenseCL	88.40	80.52	77.86	88.44	83.07	71.05
		Ours	91.68	84.64	81.25	89.27	83.77	73.68
Fundus-FFA	Linear Classification	InstDis-FFA	82.79	71.87	55.59	95.30	84.81	87.23
		InstDis-Fundus	83.22	73.65	56.16	95.88	87.34	88.89
		CMC	85.42	78.90	57.29	95.32	87.03	88.74
		SimCLR	82.06	77.94	53.00	94.85	83.54	86.02
		InstDis-All	84.22	79.42	55.93	95.32	87.34	88.25
		DenseCL	84.27	82.94	58.69	96.50	86.07	87.35
		Ours	86.46	81.43	57.80	96.32	87.61	89.09
	Finetuning	w/o pretrain	79.24	62.82	48.23	97.16	87.34	89.36
		InstDis-FFA	81.44	77.87	54.84	95.75	91.13	91.56
		InstDis-Fundus	83.41	69.13	54.79	98.06	92.40	92.85
		CMC	87.04	80.40	58.76	97.74	93.67	94.25
		SimCLR	84.78	79.89	58.86	97.68	89.87	90.69
		InstDis-All	84.13	80.64	55.66	96.46	91.13	91.56
		DenseCL	87.56	77.43	63.17	97.68	91.13	92.13
		Ours	88.58	81.42	62.50	98.59	91.35	95.23

3.1 Implementation Details

Dataset. For pretraining, we use Breast US-SWE dataset [6] and Synthesized Retinal Fundus-FFA dataset [20]. We transfer models pretrained on these two datasest to BUSI [1] and

IChallenge-PM [8] dataset respectively, to validate the transfer capacity of the method. Five-fold cross validation is conducted. Dataset details are listed in the supplementary materials.
Evaluation Protocol. We adopt linear classification and finetuning in the original and transfer dataset. Original metrics without pretraining in finetuning protocol is also reported.
Network Architecture. We adopt Vgg16 [26] as backbone network, and the projection head consists of a linear layer and ReLU, to reduce the feature dimension to 500.
Experiment Settings. All of our codes and experiments are built on PyTorch [25] with 4 NVIDIA GeForce RTX 3090 GPUs. All the input images are resized to 224×224. To keep anatomical information, we adopt relatively moderate data augmentations (flip, crop and light color jittering) in [11, 20]. In each feed forward, we set the batch size as 128. The network is optimized with SGD optimizer [2] with the learning rate of 0.03 and a weight decay of $1e-4$. We train our network for 200 epochs on Breast US-SWE dataset, and 2000 epochs (follow [20]) on Synthesized Retinal Fundus-FFA dataset.
Hyper Parameters. The temperature parameter is 0.07 and the scaling factor α in overall loss function is 3. For the feature maps, we choose the last 4 hierarchies of the feature maps and divide each feature map into 4×4 patches to calculate the consistency loss. We also analyze the sensitivity of these two parameters in Sec.3.3.

3.2 Experiment Results

To evaluate the effectiveness of our method, we compare it with some baseline models on both Breast US-SWE and Synthesized Retinal Fundus-FFA dataset.
Baseline Models. We find very few works directly working on multi-modality data. *Instance Discrimination (InstDis)* [33] is an early and essential work of contrastive learning. We carry out experiments with both single and multi-modality settings as a baseline method. *SimCLR* [6] is one of the SOTAs in SSL upon global views. *CMC* [29] focuses on contrasting images of multiple views, which is highly related to multi-modal data in our scope. *DenseCL* [31] is another SOTA method from a local and dense perspective and could capture absolute anatomical consistency in medical images. Since *SimCLR* and *DenseCL* are originally designed for uni-modality data, we consider augmented multi-modal pairs to be positive pairs to fit the original setting. The above-mentioned methods cover competitive methods in different types of self-supervised learning, providing an insight into the comparisons with representative baseline works and approaches from global, local anatomical and multi-modal viewpoints. We do not compare our method with stronger SOTAs like DINO [4] or SwAV [3], because our contributions are orthogonal to further enchance these methods.
Main Results. Table.1 (Left) shows the experimental results on two pretrained datasets. It is noticed that *InstDis* on single-modality perform relatively poorly on two datasets, because features from another modality cannot be integratedly learned in single-modality methods. Also, multi-modality methods do not always perform better. *SimCLR* achieves the worst result in the linear classification on Synthesized Retinal Fundus-FFA dataset. The main reason is that strong augmentations[6] would hurt the performance in fundus image classification. Similar phenomenon is also discovered in [20]. Moreover, *DenseCL* does not perform well, because it ignores fine-grained modality-specific attributes and incurs bias. Generally, our proposed method achieves the best performance in 9 out of 12 metrics. Especially in *AUC*, it improves on other methods by 1.57%, 0.58%, 1.04% and 1.02%. Such experimental results demonstrate the effectiveness of the propose method.
Transfer Capability. Table.1 (Right) also shows the transfer learning results of all methods. Our proposed method excels other methods in *AUC* in 3 out of 4 groups of transfer exper-

Consistency	Global	AUC	Acc	Prec
✓		75.99	54.25	49.77
	✓	87.89	81.07	79.70
✓	✓	**90.11**	**82.43**	**80.84**

Table 2: Linear classification results on US-SWE dataset to ablate each loss component. (Unit: %)

iments by 2.22%, 0.33%, 0.53% (the other one falls behind by only 0.18%). In general, transfer learning results are consistent with experimental results on pretrained datasets. Our proposed method shows superior performance over other baseline models, which indicates that our method could generalize to different downstream datasets.

3.3 Ablation Study

Analysis of Consistency Loss and Global Contrastive Loss To validate the effectiveness of two loss terms, we train our unsupervised model with global loss and consistency loss separately. As is shown in Table.2, the overall loss function $L = L_C + \alpha L_{Glo}$ achieves the best performance over the other two. When trained with consistency loss alone, the model performs poorly. This is because without global loss, the global representation is not optimized to capture the instance and modality level relationship, which is important for semantic downstream tasks (i.e. classification). Moreover, Fig.3 indicates that the global loss and consistency loss would converge to smaller values when they are trained together versus when they are trained individually. Therefore, the two loss functions are mutually beneficial. Notably, with the contribution of consistency loss, smaller global contrastive loss denotes that positive pairs are closer in feature space. This is further validated in feature space in Fig.4. Corresponding image pairs (US and SWE modality) cluster more closely, when the two loss functions are trained unitedly. In contrast, the two modalities in each pair get scattered when trained with global contrastive loss alone. Generally, two loss terms are helpful to each other, and consistency loss designed to align similarity distribution for anatomical consistency promotes global representation.

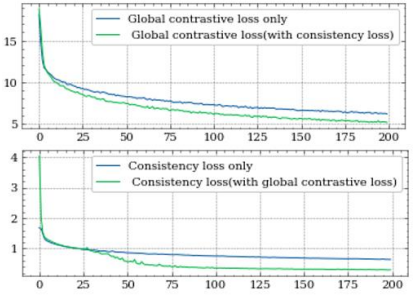


Figure 3: Global contrastive loss and consistency loss during training, when trained alone and unitedly on Breast US-SWE dataset.(Top: global contrastive loss; Bottom: consistency loss)

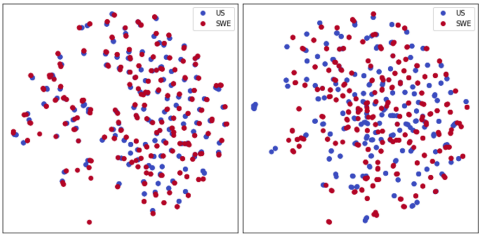


Figure 4: A t-SNE [30] Visualization of learned feature embedding of US and SWE modalities (Left: Trained with overall loss. Right: Trained with only global contrastive loss. One fifth samples from training data.).

Analysis of Details in Anatomical Consistency For the input of the branch, we only calcu-

Table 3: Experiment results on different technical details on Breast US-SWE dataset in linear classification (Unit:%). Each group of experiments studies a component in the branch of anatomical consistency, and they are compared with the adopted settings in bold.

No.	Input	Consistency	Hierarchy	Patch Num	AUC	Acc	Prec
1	With Aug	Correlation Matrix	[1,2,3,4]	(4,4)	89.19	81.97	80.65
2	Aligned	Normed Correlation	[1,2,3,4]	(4,4)	87.62	80.99	79.38
	Aligned	Local Contrast	[1,2,3,4]	(4,4)	88.70	81.50	81.94
	Aligned	KL	[1,2,3,4]	(4,4)	88.50	81.79	81.22
3	Aligned	Correlation Matrix	[4]	(4,4)	89.90	83.12	80.05
	Aligned	Correlation Matrix	[3,4]	(4,4)	89.24	81.68	77.42
	Aligned	Correlation Matrix	[2,3,4]	(4,4)	90.46	82.16	78.79
4	Aligned	Correlation Matrix	[1,2,3,4]	(8,8)	89.48	81.62	79.83
	Aligned	Correlation Matrix	[1,2,3,4]	(8,8)&(4,4)*	89.82	82.16	78.79
Adopted	**Aligned**	**Correlation Matrix**	**[1,2,3,4]**	**(4,4)**	**90.11**	**82.43**	**80.84**

$*$ $(8,8)$ applied in the first two hierarchies of feature maps, and $(4,4)$ applied in the last two. Feature maps from the first two hierarchies are of greater sizes, thus it is natural to divide them into more patches.

late the consistency loss of ***Aligned*** feature maps without data augmentations, so we study the impact of augmentations (***With Aug***). For the calculation of consistency loss, there are some other alternatives. ***Normed Corr*** denotes that we normalize the similarity distribution with softmax. ***Local Contrast*** performs contrastive learning on larger patches rather than pixels in [41] for anatomical consistency, which is another alternative to give more tolerance to fine-grained modality-specific attributes. ***KL*** means that we adopt KL-divergence to align correlation matrix. Moreover, we also investigate some hyper parameters in this branch: hierarchies of feature maps utilized in consistency loss and the number of output patches.

1) Group No.1: Adding augmented feature maps would influence model performance (decrease by 0.92% in *AUC*). Augmentations would shuffle the corresponding patches, thus spatial similarity distribution across modalities cannot be strictly aligned.

2) Group No.2: It first shows that normalization harms model performance, because softmax normalization would smooth the similarity distribution, and the model may neglect some slight similarity difference which might be non-negligible. Moreover, ***Local Contrast*** could not achieve as good performance as ours. First of all, working on larger patches does not pay much attention to modality-specific attributes fundamentally. Secondly, it would only focus on discriminating a single patch from others, neglecting the general structure of patch-wise relationships, while our proposed method could capture overall similarity distribution within the entire image. We also observe that KL divergence does not perform well to align the correlation matrix. We infer that the normalization during calculating KL divergence accounts for its poor performance.

3) Group No.3: It investigates the sensitivity of feature map hierarchies. We take the fourth (also the last) layer of the feature maps as a requirement, and investigate how different hierarchies influence model performance. It is noted that the consistency loss propagates back to the neural network from the layer where it is produced. To ensure the backbone network is fully optimized, the last layer of the feature maps should be included. Table.3 shows that with the fourth layer included in the hierarchy of the feature maps, changing hierarchies would not significantly influence the overall performance.

4) Group No.4: For the number of output patches in feature maps, we mainly conduct experiments with a patch number of 4×4 and 8×8. Since the computational complexity of

similarity distribution is $O(n^4)$ for $n \times n$ output patches, we do not consider dividing the feature maps into more patches for simplicity. We observe that the number of output patches in feature maps has little influence on the overall performance of the model.

3.4 Qualitative Results

To better demonstrate the effectiveness of the proposed method, we visualize the similarity distribution in Fig.5. Given the same anchor(yellow), our proposed method captures better anatomy consistency, since it obtains more consistent similarity distribution across modals.

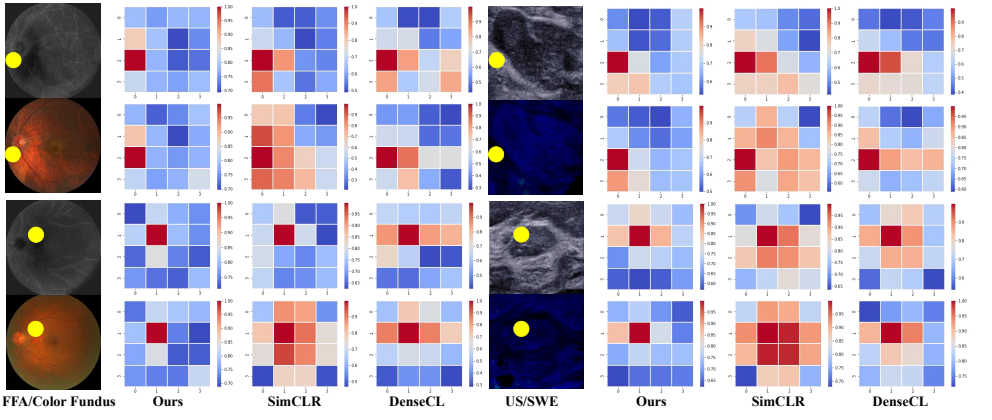


Figure 5: Visualization of similarity distribution (4×4 local patches) of the given anchor.

4 Discussion

Our proposed method mainly focuses on aligned multi-modality medical data, while unaligned modalities are more easily accessible in some cases. For such a more challenging problem, we could extend our work from different aspects, for example, 1) Applying self-supervised image registration as pre-processing at input- or feature-level. 2) Using robust contrastive learning loss in similarity distribution to relieve the noise caused by the unaligned features. These improvements will be our future work.

5 Conclusion

In this paper, we present a novel anatomy-aware self-supervised learning method among aligned multi-modality data. Our key idea is to capture anatomical consistency across modalities, with tolerance to modality-specific attributes. Our proposed method achieves this goal by constructing a correlation matrix to represent similarity distribution and designing a consistency loss to align the distribution. Global-invariant features are also learned in a contrastive manner. Extensive experimental results demonstrate that our method achieves superior performance to previous methods. Detailed ablation studies also validate the effectiveness of aligning similarity distribution for anatomical consistency. Future work could be extended to more general cases to unaligned medical images or natural images.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China 62171282, 111 project BP0719010, STCSM 18DZ2270700, Shanghai Jiao Tong University Science and Technology Innovation Special Fund ZH2018ZDA17, and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409.

[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[5] Kun Chen, Yuanfan Guo, Canqian Yang, Yi Xu, Rui Zhang, Chunxiao Li, and Rong Wu. Enhanced breast lesion classification via knowledge guided cross-modal and semantic data augmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 53–63. Springer, 2021.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[8] Huazhu Fu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE transactions on medical imaging*, 37(7):1597–1605, 2018.

[9] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[10] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):162–169, 2019.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[12] Álvaro S Hervella, José Rouco, Jorge Novo, and Marcos Ortega. Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis. *Expert Systems with Applications*, 185:115598, 2021.

[13] Álvaro S Hervella, José Rouco, Jorge Novo, and Marcos Ortega. Multimodal image encoding pre-training for diabetic retinopathy grading. *Computers in Biology and Medicine*, 143:105302, 2022.

[14] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[15] Olle G Holmberg, Niklas D Köhler, Thiago Martins, Jakob Siedlecki, Tina Herold, Leonie Keidel, Ben Asani, Johannes Schiefelbein, Siegfried Priglinger, Karsten U Kortuem, et al. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. *Nature Machine Intelligence*, 2(11):719–726, 2020.

[16] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9 (1):2, 2020.

[17] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Learning image representations by completing damaged jigsaw puzzles. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 793–802. IEEE, 2018.

[18] Jong-Ha Lee, Yeong Kyeong Seong, Chu-Ho Chang, Jinman Park, Moonho Park, Kyoung-Gu Woo, and Eun Young Ko. Fourier-based shape feature extraction technique for computer-aided b-mode ultrasound diagnosis of breast tumor. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6551–6554. IEEE, 2012.

[19] Sook Sam Leong, Jeannie Hsiu Ding Wong, Mohammad Nazri Md Shah, Anushya Vijayananthan, Maisarah Jalalonmuhali, Nur Hidayati Mohd Sharif, Nurul Khairyah Abas, and Kwan Hoong Ng. Stiffness and anisotropy effect on shear wave elastography: a phantom and in vivo renal study. *Ultrasound in medicine & biology*, 46(1): 34–45, 2020.

[20] Xiaomeng Li, Mengyu Jia, Md Tauhidul Islam, Lequan Yu, and Lei Xing. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 39(12):4023–4033, 2020.

[21] Xirong Li, Yang Zhou, Jie Wang, Hailan Lin, Jianchun Zhao, Dayong Ding, Weihong Yu, and Youxin Chen. Multi-modal multi-instance learning for retinal disease recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2474–2482, 2021.

[22] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[23] Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020.

[24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] Qiuchang Sun, Xiaona Lin, Yuanshen Zhao, Ling Li, Kai Yan, Dong Liang, Desheng Sun, and Zhi-Cheng Li. Deep learning vs. radiomics for predicting axillary lymph node metastasis of breast cancer using ultrasound images: don't forget the peritumoral region. *Frontiers in oncology*, 10:53, 2020.

[28] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis. In *International Conference on Information Processing in Medical Imaging*, pages 661–673. Springer, 2021.

[29] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[30] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[31] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

[32] Zhaoqing Wang, Qiang Li, Guoxin Zhang, Pengfei Wan, Wen Zheng, Nannan Wang, Mingming Gong, and Tongliang Liu. Exploring set similarity for dense self-supervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16590–16599, 2022.

[33] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[34] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

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

[36] Yao Zhang, Jiawei Yang, Jiang Tian, Zhongchao Shi, Cheng Zhong, Yang Zhang, and Zhiqiang He. Modality-aware mutual learning for multi-modal medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 589–599. Springer, 2021.