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

Hongyu Hu\textsuperscript{1}\hfill \hspace{1cm}\textsuperscript{1}MoE Key Lab of Artificial Intelligence  
Shanghai Jiao Tong University  
Shanghai, China  
Tiancheng Lin\textsuperscript{1}\hfill \textsuperscript{1}ltc19940819@sjtu.edu.cn  
Yuanfan Guo\textsuperscript{1}\hfill \textsuperscript{1}gyfastas@sjtu.edu.cn  
Chunxiao Li\textsuperscript{2}\hfill \textsuperscript{2}lcx198910@126.com  
Rong Wu\textsuperscript{2}\hfill \textsuperscript{2}wurong7111@163.com  
Yi Xu\textsuperscript{*1}\hfill \textsuperscript{1}xuyi@sjtu.edu.cn

\textsuperscript{1}MoE Key Lab of Artificial Intelligence  
Shanghai Jiao Tong University  
Shanghai, China  
\textsuperscript{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 \cite{16, 22}, such as context-instance relationships \cite{7, 9, 17}, instance-instance contrast \cite{6, 11, 14, 24, 33}, or dense contrast \cite{23, 31, 34}. This characteristic is of great help to the field of medical

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.
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-
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.

A 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:

$$M = \{(m_1^A, m_1^B), (m_2^A, m_2^B), \ldots, (m_N^A, m_N^B)\}; \hat{M} = \{(\hat{m}_1^A, \hat{m}_1^B), (\hat{m}_2^A, \hat{m}_2^B), \ldots, (\hat{m}_N^A, \hat{m}_N^B)\}$$

The neural network $G_\theta$ consists of $\ell$ stacked convolutional layers $\theta$ as the backbone, followed by a projection head $\Theta$:

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

Our goal is to learn a good feature embedding network $G_\theta$ in an unsupervised manner, which can embed image $m_1^A$ and $m_1^B$ into highly distinguishable vectors $f_1^A$ and $f_1^B \in \mathbb{R}^d$, 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 \(G_{\theta}\), for each convolutional layer \(\theta\), 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 \(\lceil \frac{H}{N_h} \rceil \times \lceil \frac{W}{N_w} \rceil\). The embedded feature \(s_i\) of the patch \(p_i\) is defined as:

\[
s_i = \frac{\sum_{(a,b) \in p_i} v(a,b)}{\lVert \sum_{(a,b) \in p_i} v(a,b) \rVert_2} \tag{3}
\]

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.

\[
A_{i,j} = \text{sim}(p_i, p_j) = s_i^T s_j, \ (i, j \in N_h \times N_w) \tag{4}
\]
where \( \text{sim}(\cdot) \) 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:

\[
L_{SD} = \frac{1}{(N^hN^w)^2} \sum_j \sum_i (A_{i,j}^{(M_A)} - A_{i,j}^{(M_B)})^2
\]

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

\[
L_C = \frac{1}{|Q|} \sum_{\theta_i \in Q} L_{SD}^{(\theta_i)}
\]

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 \( i \)th layer.

### 2.4 Global-Invariant Feature Representation

After the projection head \( \Theta \), the network \( G_\theta \) embeds the input data \((m_A, m_B)\) and \((\hat{m}_A, \hat{m}_B)\) into high-dimensional feature vectors \((f_A, f_B)\) and \((\hat{f}_A, \hat{f}_B)\). We then normalize all the feature vectors by \( l_2 \) normalization, i.e., \( \|f_A\|_2 = \|f_B\|_2 = \|\hat{f}_A\|_2 = \|\hat{f}_B\|_2 = 1 \). 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:

\[
L_{AUG}^{M_A} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(f_{A}^i \cdot \hat{f}_{A}^i / \tau)}{\sum_{j=1}^{N} \exp(f_{A}^i \cdot \hat{f}_{A}^j / \tau)}; \quad L_{AUG}^{M_B} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(f_{B}^i \cdot \hat{f}_{B}^i / \tau)}{\sum_{j=1}^{N} \exp(f_{B}^i \cdot \hat{f}_{B}^j / \tau)}
\]

\[
L_{AUG} = L_{AUG}^{M_A} + L_{AUG}^{M_B}
\]

where \( N \) denotes the size of the sample batch, dot product \( \cdot \) is used to calculate cosine similarity and \( \tau \) 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:

\[
L_{M \rightarrow B} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(f_{A}^i \cdot f_{B}^i / \tau)}{\sum_{j=1}^{N} \exp(f_{A}^i \cdot f_{B}^j / \tau)}; \quad L_{M \rightarrow A} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(f_{B}^i \cdot f_{A}^i / \tau)}{\sum_{j=1}^{N} \exp(f_{B}^i \cdot f_{A}^j / \tau)}
\]
\[ L_M = L_M^{A\rightarrow B} + L_M^{B\rightarrow A} \quad (10) \]

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

\[ L_{Glo} = L_{AUG} + L_M \quad (11) \]

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

\[ L = L_C + \alpha L_{Glo} \quad (12) \]

where \( \alpha \) 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:%)

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

### 3.1 Implementation Details

**Dataset.** For pretraining, we use Breast US-SWE dataset [5] and Synthesized Retinal Fundus-FFA dataset [20]. We transfer models pretrained on these two datasets to BUSI [6] and
In Challenge-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 $224 \times 224$. 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 $1 \times 10^{-4}$. 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 $\alpha$ 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 $4 \times 4$ 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 enhance 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 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-
ments 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.

<table>
<thead>
<tr>
<th>Consistency</th>
<th>Global</th>
<th>AUC</th>
<th>Acc</th>
<th>Prec</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>75.99</td>
<td>54.25</td>
<td>49.77</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>87.89</td>
<td>81.07</td>
<td>79.70</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>90.11</td>
<td>82.43</td>
<td>80.84</td>
</tr>
</tbody>
</table>

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

![Figure 3: Global contrastive loss and consistency loss during training, when trained alone and unitedly on Breast US-SWE dataset.](image)

![Figure 4: A t-SNE Visualization of learned feature embedding of US and SWE modalities](image)

**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.

<table>
<thead>
<tr>
<th>No.</th>
<th>Input</th>
<th>Consistency</th>
<th>Hierarchy</th>
<th>Patch Num</th>
<th>AUC</th>
<th>Acc</th>
<th>Prec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>With Aug</td>
<td>Correlation Matrix</td>
<td>[1,2,3,4]</td>
<td>(4,4)</td>
<td>89.19</td>
<td>81.97</td>
<td>80.65</td>
</tr>
<tr>
<td>2</td>
<td>Aligned</td>
<td>Normed Correlation</td>
<td>[1,2,3,4]</td>
<td>(4,4)</td>
<td>87.62</td>
<td>80.99</td>
<td>79.38</td>
</tr>
<tr>
<td></td>
<td>Aligned</td>
<td>Local Contrast</td>
<td>[1,2,3,4]</td>
<td>(4,4)</td>
<td>88.70</td>
<td>81.50</td>
<td>81.94</td>
</tr>
<tr>
<td></td>
<td>Aligned</td>
<td>KL</td>
<td>[1,2,3,4]</td>
<td>(4,4)</td>
<td>88.50</td>
<td>81.79</td>
<td>81.22</td>
</tr>
<tr>
<td>3</td>
<td>Aligned</td>
<td>Correlation Matrix</td>
<td>[1,2,3,4]</td>
<td>(8,8)</td>
<td>89.48</td>
<td>81.62</td>
<td>79.83</td>
</tr>
<tr>
<td></td>
<td>Aligned</td>
<td>Correlation Matrix</td>
<td>[1,2,3,4]</td>
<td>(8,8)&amp;(4,4)</td>
<td>89.82</td>
<td>82.16</td>
<td>79.79</td>
</tr>
<tr>
<td></td>
<td>Adopted</td>
<td>Aligned</td>
<td>Correlation Matrix</td>
<td>[1,2,3,4]</td>
<td>(4,4)</td>
<td>90.11</td>
<td>82.43</td>
</tr>
</tbody>
</table>

* (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 [31] 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.](image-url)

### 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


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.


[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.


