几何视觉相似性学习用于三维医学图像自监督预训练

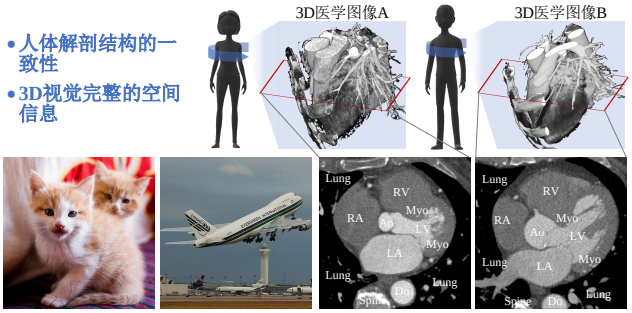
何宇霆 [1], 杨冠羽 [1]*, 葛荣骏 [2], 陈阳 [1], Jean-Louis Coatrieux[3], 王柏予 [4], 李硕 [5]
[1] 东南大学 [2] 南京航空航天大学
[3]University of Rennes 1 [4]Western University [5]Case Western Reserve University

Abstract

学习图像间的相似性对于三维医学图像的自监督预训练至关重要，因为它们共享大量相同的语义区域。然而，度量过程中语义先验的缺乏和 *3D* 医学图像中语义无关的变化使得难以可靠度量图像间相似性，阻碍了对相同语义特征间的一致性的学习。我们研究了这项任务中具有挑战性的问题，即学习图像之间的一致表征，以获得相同语义特征的聚类效应。我们提出了一种新的视觉相似性学习范式——几何视觉相似性学习，它将人体间拓扑不变性的先验嵌入到图像间相似性的度量中，以实现语义区域的一致表示。为了驱动这种范式的学习，我们进一步构建了一种新的几何匹配头，*Z* 形匹配模块，以同时学习语义区域的全局和局部相似性，引导不同粒度的图像间语义特征的高效学习。我们的实验表明，在四个具有挑战性的 *3D* 医学图像任务上，利用我们的图像间相似性学习的预训练产生了更强大的场景内、场景间和全局-局部的迁移能力。我们的代码和预训练模型在*https://github.com/YutingHe-list/GVSL*上公开.

1. 介绍

学习图像间相似性 [24, 31, 42] 对于三维医学图像（如 CT、MRI）的自监督预训练（self-supervised pre-training，SSP）[19] 至关重要。如图1所示，与在 SSP 中被广泛研究的自然图像不同，由于人体解剖结构的一致性 [26] 和 3D 视觉中完整的空间信息 [33]，三维医学图像间共享大量相同的语义区域，为有效的 SSP 带来了强有力的先验。因此，其目标在于约束预训练网络



a）自然图像间有着极大的语义差异 b）3D医学图间共享大量的语义区域

图 1. 三维医学图像与自然图像的差异图。a）自然图像之间相似性较弱，具有较大的语义差异。b）由于人体解剖结构的一致性和空间信息的完备性，三维医学图像之间共享大量相同的语义区域，图像间相似性大。

以在没有注释的图像之间一致地表示这些相同的语义区域。这将对相同的语义特征带来强大的聚类效应，使得预训练网络具有良好的表征能力，最终能够进行有效地迁移学习潜在的下游具体医学图像分割任务。因此，它将有效地训练数据不足的巨大挑战 [43]，即使在目标任务数据量较少的情况下依旧能够获得更强大的模型。

虽然现有的 SSP 的相关研究工作已经在其各自的目标任务中取得了良好的效果，但在三维医学图像的图像间相似性学习方面仍存在一定的局限性。*1*）基于聚类的方法 [3, 22] 根据图像在特征空间中的聚类模式来度量图像间特征的相似程度，并学习聚合相同潜在类别的特征，分离不同潜在类别的特征。然而，基于聚类的 SSP 方法通常采用马氏距离或欧氏距离 [2] 作为距离度量函数，使距离度量受到图像内语义无关变化（如外观变化）的干扰（图2），大大限制了对潜在类别判断的准确性，限制最终的自监督学习。*2*）基于对比

*通讯作者: yang.list@seu.edu.cn

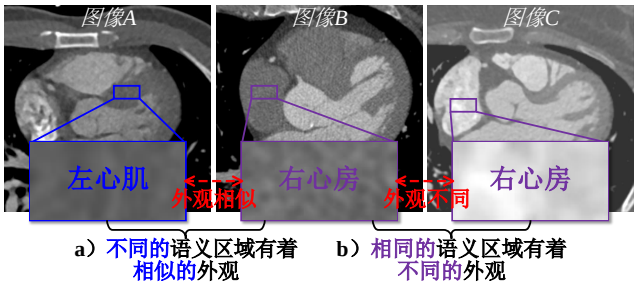


图 2. 度量图像间的相似性具有极大的挑战。a）图 A 的左心肌和图 B 的右心房之间存在着很大的相似性。b）由于采集仪器扫描协议的变化，在图 B 和图 C 的两个右心房之间存在着很大的外观差异。

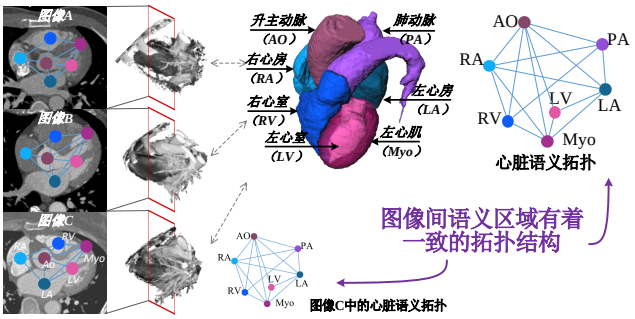


图 3. 三维医学图像之间语义区域的拓扑不变性提供了一种构建可靠的图像间相似性度量的潜在方法。

学习的方法 [4,5] 直接通过对比来分离来自不同图像的特征，试图学习图像间的特征相异性。但是过度学习图像间的不同将使得网络容易把相同的语义区域表征为不同的特征，使得模型无法学习图像间的相似性。虽然其他一些对比学习的研究 [5,8,39] 通过去除了对比学习中的负对学习来避免这种情况，仅依靠学习同一图像变换前后的一致性来驱动 SSP，但使这些方法仍然无法学习图像间相同语义的一致性。*3）基于生成的方法* [21,23,38,43] 通过人工设计的变换方法（例如，旋转 [21]）来生成代理标签（pretext label），然后训练网络来预测这些代理标签以驱动 SSP。但是这些通过人工设计的变换方法隐式地将人的偏好强加到 SSP 学习中，使得预训练的网络参数关注于代理标签所偏好特征，导致获得的预训练模型对场景变化敏感 [23]。例如，旋转生成的代理标签 [21] 将使网络偏向于图像中对象的位置和姿态特征，当某些场景中对象的位置和姿态是一种与语义无关的信息时，这些图像将被错误表征。

让我们重新思考上述已有工作的局限性，*大范围图像间相似性的错误度量是三维医学图像 SSP 面临的关键挑战。*该挑战干扰了在自监督学习过程中对潜在相同语义对象的发现，阻碍了模型对相同语义区域学习一致的表征。如图2所示，与语义无关的图像特征变化使得三维医学图像间具有不同的外观，导致不同的语义区域具有相似的外观（图 A 和 B 之间的左心肌和右心房），而相同的语义区域在图像之间具有不同的外观（图 B 和 C 之间的右心房）。类似于基于聚类 [3,22] 这种利用欧氏距离或马氏距离在特征空间中的直接相似性度量，由于其度量过程中缺少语义相关的先验知识而变得敏感，容易受到语义无关变化的干扰。因此，在无监督学习的情况下，一旦被提取的特征受到外观干扰发生变化，就会导致大规模的语义相似性的错误测量，从而错误地聚合或分离语义对象，最终使得预训练模型学习到的表示偏离真实的语义分布。

三维医学图像中语义区域的拓扑不变性 [25] 为 SSP 提供了一种构建可靠的图像间相似性度量的潜在方法。如图3所示，由于人体内部解剖结构的一致性 [26]，三维医学图像的语义区域之间具有一致的上下文拓扑（例如人体心脏的四个腔室具有固定的上下空间关系），并且不同图像中的相同语义区域在也具有相似的形状（例如图中的主动脉具有稳定的管状结构），因此图像中的语义区域之间能够构建稳定的拓扑结构，形成了一种可靠的先验知识。根据该语义区域拓扑不变性先验，我们能够将图像间的区域经由某种拓扑不变的映射 [11] 在空间中对齐，从而即使图像在外观上有大的变化也能够可靠地构建特征间的对应关系。一种直观的策略是使用配准（Registration）或几何匹配（Geometric Matching, GM）方法 [12,13,15,16,30] 来发现图像像素间的对应关系，并使用这些对应索引来约束网络输出对应像素位置区域的特征一致。然而，当配准或几何匹配出现错误时，这些索引的误差会带来误的对应关系，学习到错误的表征。在本章中我们提出了高效三位医学图像自监督表征预训练的新型假说: *三维医学图像之间语义区域的拓扑不变性先验能够驱动模型发现图像间特征对应关系，从而引导网络可靠学习图像间的相似性。*

在本章中，我们提出了一种新的自监督表征预训练学习方法，几何视觉相似性学习（Geometric Visual Similarity Learning，GVSL），用于可靠学习三维医学图像间的相似性。该算法将拓扑不变性先验嵌入到相

似性度量和学习过程中，训练网络利用从两张三维医学图像上分别提取的特征来预测语义区域的对应关系，即配准，从而驱动学习图像间的一致性表征。由于拓扑不变性这种有效的先验知识的帮助，图像间相似性将在医学图像中语义区域固有的拓扑结构下被度量，从而避免了与语义无关的外观变化的干扰。当模型学习更准确地估计对应关系时，反向传播的梯度将训练神经网络在特征空间中聚合对应的特征以获得更一致的表示。为了有效地同时学习全局和局部的表征，我们进一步提出了新型的自监督学习模块，Z 形匹配模块。该学习模块通过同时学习仿射变换（全局配准）和弹性形变（局部配准）[13] 来驱动模型同时学习图像间全局和局部相似性，从而使网络拥有同时高效学习全局和局部表征的能力，为对不同尺度目标结构的医学图像分割任务都提供强大的迁移能力。

本章的贡献总结如下：1）将图像间的相似性学习应用于三维医学图像自监督表示学习中，从而预训练神经网络以学习语义一致的表示，在无需标注的情况下，使模型拥有强大迁移能力的预训练参数。2）提出了几何视觉相似性学习（GVSL），该方法将三维医学图像的拓扑不变性先验嵌入到相似性度量中，从而可靠地度量图像间的相似性，促进模型学习图像间相同语义区域的一致表示。3）提出了一种新型的自监督学习模块，Z 形匹配模块，该模块同时学习仿射变换和弹性形变配准，同时学习强大的全局和局部表示，最终获得了对不同语义粒度表示的有效优化，为不同尺度目标结构的医学图像分割任务提供了强大的迁移学习能力。

2. 相关工作

自监督预训练中的相似性学习：学习相似性 [24,41] 的目标是学习图像中相似的对象获得一致的表示和学习相异的对象获得有鉴别力的表示，作为视觉自监督表征学习的一个基本任务 [19]，正在极大地提升模型在学习下游任务时初始化的表征能力。正如我们已经在第1节所介绍的，学习图像间的相似性有三个主要范式。对比学习 [4,5,14,37] 范式认为同一图像的不同数据增强视图是相似的对象，而不同图像为相异的对象，从而学习他们的一致性和鉴别性来获得有效表征。但是，这类方法无法学习图像间的相似性，简单地认为来自不同图像的特征为相异对象，从而分离这些特征将极大地干扰三维医学图像的表示。基于聚类的方法 [3,22] 利用度量的方法来获得相似与相异对象簇，从而聚集相似对象以学习图像间的相似性，这类方法很容易发生聚类错误的问题，从而将本身不同类别的图像认为是相似对象而聚集，相似性干扰了模型学习有效的表征。基于生成的方法 [9,10,21,23,38,43] 通过人工设计获得生成的代理标签，并约束网络预测这些标签。然而，它隐含地嵌入了手动设计的偏见，使网络忽略一些潜在的方面，并限制了在某些特定情况下的迁移。

几何匹配与配准：配准（或称为几何匹配）[12,13,16,17,30,34] 将图像的语义区域对齐到同一空间坐标系中，从而提供两个图像之间的对应索引。它具有两个级别的变换：1）仿射变换 [13,40] 能够在全局上对齐图像分布。它计算一个由旋转、缩放、平移和剪切操作组成的变换矩阵，将图像转换为在全局上对齐的形式。2）弹性形变 [13,15,16,34] 能够在局部对齐图像细节。它预测一个体素级位移矢量场（Displacement vector field, DVF），获得图像间体素的对应关系，并通过空间变换操作 [18] 将图像局部对齐。最近，由于深度学习的发展，基于深度学习的配准 [13,16,30] 利用学习的方式获得端到端的对应预测能力，为我们的框架设计提供了一种潜在的方案。

3. 方法

GVSL 框架（图4）能够从无标签的三维医学图像中学习具有图像间相似性的通用视觉表征。它将拓扑不变性先验嵌入到相似度度量过程中，从而驱动模型学习可靠的相似语义区域的一致表征。

3.1. GVSL 用于学习图像间相似性

如图4 a）所示，GVSL 将学习图像间相似性的过程建模为网络所表征的特征中预测图像间配准的过程，嵌入图像中语义区域的拓扑不变性先验，从而在反向传播中训练网络来为相同的语义区域表征一致的特征。

3.1.1 GVSL 的方法细节描述 GVSL 的目标是学习一个相同语义区域的特征在不同的图像间也具有的强大聚类效应的通用网络表征。如图4所示，它使用两个共享权重的神经网络 N_θ 来表征来自两个图像 x_A, x_B 的特征 f_A, f_B，其中 θ 是网络内部的可学习权重。这些特征被进一步输入配准头 G_ξ（即，我们框架中的 Z 形匹配模块，Sec.3.2），以学习图像之间语义区域的对应关系，从而使 N_θ 获得这些相同语义特征的一致表示。

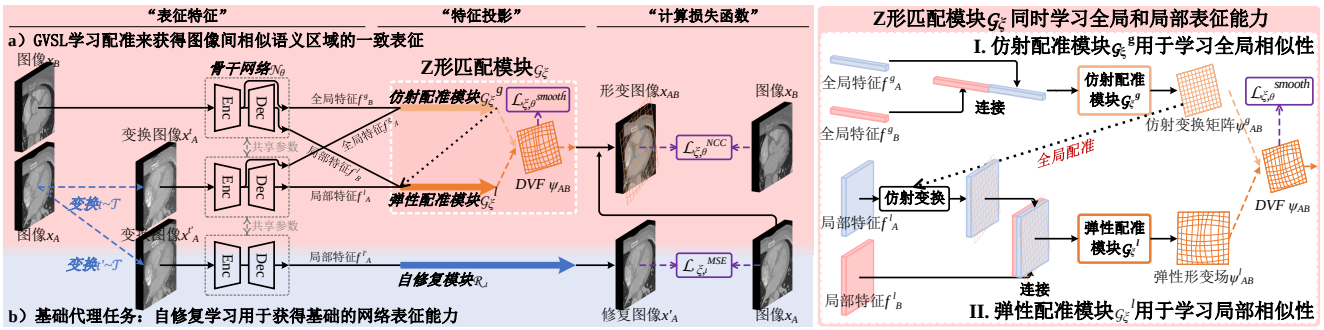


图 4. GVSL 框架：a）GVSL 从图像的特征中学习配准，通过反向传播的梯度驱动图像之间的相似性学习。b）为了有效学习，在学习配准的同时，它还进一步融合了一个基础的自我修复任务，为网络的基础表征能力提供了一种预热，从而更加高效地驱动配准学习。

给定一组三维医学图像 $\mathbb{D}_{train}$，从 $\mathbb{D}_{train}$ 中均匀随机采样两个图像 $\{x_A, x_B\} \sim \mathbb{D}_{train}$（$A$ 和 B 代表不同的图像），并从一个变换集 T 中采样一个变换操作 $t \sim T$。在本研究中，我们使用修补、体素局部抖动和图像强度非线性作为变换集 T。GVSL 通过将变换 t 作用于图像上，使 x_A 变换为一个变化视图 $x^t_A \triangleq t(x_A)$，从而提高了训练图像的多样性。网络输出 x^t_A 和 x_B 的全局和局部特征 $\{f^g_A, f^l_A\} \triangleq N_\theta(x^t_A), \{f^g_B, f^l_B\} \triangleq N_\theta(x_B)$。接着这些特征输入配准头 G_ξ 以输出一个位移向量场 (DVF)$\psi_{AB} \triangleq G_\xi(f^l_A, f^l_B, f^g_A, f^g_B)$，该向量场中的每个向量对应于移动图像中的每个像素移动到固定图像空间中的位置，即是两个图像之间体素的对应关系。最后通过空间变换，将图像 x_A 变换到图像 x_B 以配准两个图像，得到一个被几何匹配的形变图像 $x_{AB} \triangleq \psi_{AB}(x_A)$。其中，空间变换操作我们采用了 [18] 的实现方式，即对于图像 x_A 中的每个体素 p，位移向量场 ψ_{AB} 将 p 位移至图像空间中的一个新的（亚像素）体素位置 $\psi_{AB}(x_A(p))$。然后，线性插值该位移后在亚像素位置上的体素到八个相邻体素的整数位置。该过程表示为：$\psi_{AB}(x_A(p)) = \Sigma_{q\in\psi_{AB}(\mathbb{Z}(p))} x_A(q)\Pi_{d\in x,y,z}(1 - |\psi_d(x_A(p)) - q_d|)$，其中，$\psi_{AB}(\mathbb{Z}(p))$ 是 $\psi(x_A(p))$ 的相邻体素，x, y, z 是三维图像的 x, y, z 轴。

我们利用归一化互相关（NCC）来评估两个图像之间的对齐程度，从而间接评估所预测的对应关系（DVF）的准确性。我们训练框架来最小化这个损失函数 L^{NCC}，从而驱动模型优化这个对应关系以获得更好的配准：

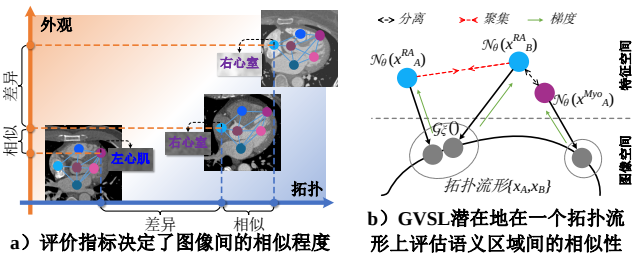
$$\mathcal{L}^{NCC}_{\theta,\xi} \triangleq \sum_{p\in\Omega} \frac{(\sum_{p_i}(x_B(p_i) - \hat{x_B}(p))(x_{AB}(p_i) - \hat{x_{AB}}(p)))^2}{(\sum_{p_i}(x_B(p_i) - \hat{x_B}(p))^2)(\sum_{p_i}(x_{AB}(p_i) - \hat{x_{AB}}(p))^2)}, \tag{1}$$

其中 p 是图像空间 Ω 中体素的位置，i 是 p 的索引。$\hat{x}_B(p)$ 是图像强度的局部均值：$\hat{x}_B(p) = \frac{1}{n^3}\sum_{p_i} x_B(p_i)$（$\hat{x}_{AB}(p)$ 与其计算方式相同）。为了保持图像配准前后的拓扑不变性，我们进一步在 DVF 上计算平滑损失 L^{smooth}，以训练网络在保持拓扑一致的条件下学习语义区域的对应关系：

$$L^{smooth}_{\theta,\xi} \triangleq \sum_{p\in\Omega} \| \bigtriangledown \psi_{AB}(p)\|^2. \tag{2}$$

在每个训练步骤中，我们使用梯度进行反向传播，通过最小化 $L^{GVSL}_{\theta,\xi} = L^{NCC}_{\theta,\xi} + L^{smooth}_{\theta,\xi}$ 来优化网络权重 θ 和配准头的权重 ξ。因此，框架的总优化过程可被描述为一个基于梯度下降的过程 $\theta, \xi \leftarrow optimizer(\theta, \xi, \nabla_{\theta,\xi} L^{GVSL}_{\theta,\xi}, \eta)$，其中 $optimizer$ 是训练中所使用的优化器，η 是学习率。

3.1.2 GVSL 背后的算法直觉 GVSL 中配准的学习过程、相似性损失 L^{NCC} 和用于拓扑保持的平滑损失 L^{smooth} 共同组成的模型学习过程是一种带有拓扑不变先验的隐式度量 [1]。如图5 a）所示，相似性度量的依据决定了图像之间的相似度大小。由于外观上与语义无关的变换，仅使用图像外观的相似性将大大限制度量的可靠性。而配准约束度量相似性的过程在拓扑不变条件下进行，避免了由于外观造成的误度量（如图2所示）。我们可以将其描述为在 $\{x_A, x_B\}$ 的流行平



a）评价指标决定了图像间的相似程度　b）GVSL潜在地在一个拓扑流形上评估语义区域间的相似性

图 5. GVSL 背后的算法直觉：配准中的拓扑不变性先验将拓扑流形嵌入到度量过程中，从而为图像之间的相似性提供了有效的度量，从而反向通过梯度引导模型学习相同语义的特征拥有一致的表征。

面上进行的相似性度量：

$$\min_{\theta,\xi} L(G_\xi(N_\theta(x_A), N_\theta(x_B)); \{x_A, x_B\}) \qquad (3)$$

如图 5 b）和式 3所示，它隐式地将图像 x_A, x_B 中语义结构内在拓扑流形嵌入到相似性度量过程中（L^{NCC}），从而获得更加可靠的相似性评价。x_A^{RA}，x_B^{RA} 和 x_A^{Myo} 分别为图像 x_A 和 x_B 上的右心房区域和图像 x_A 上的左心肌区域。由于图像 x_B 的右心房和图像 x_A 的左心肌区域在外观上相似，和图像 x_A 右心房外观上不同，在特征空间中，两个不同语义的特征将更加接近而相同语义的特征将更远。因此某些直接的特征距离度量，如常用的欧氏距离和马氏距离，将极容易发生错误，最终导致错误的表征学习。而我们的 GVSL 使用配准头 G_ξ 将特征空间中的特征（$N_\theta(x_A^{RA})$、$N_\theta(x_B^{RA})$、$N_\theta(x_A^{Myo})$）进一步映射到图像空间中的拓扑流形上。因此，由于有效的拓扑不变性先验，相同语义区域的特征之间的距离将变得更近，从而通过梯度来学习有效的图像相似性。

在整个训练过程中，网络 N_θ 中的特征表征学习和配准头 G_ξ 中的语义区域对应学习是一个博弈关系[32]。配准头 G_ξ 学习从表征的特征 f_A, f_B 中预测图像空间中语义区域的对应关系。而网络 N_θ 学习向配准头 G_ξ 提供语义区域的特征，来进行区域对应的学习。而为了获得更精确的对应关系，配准头也将通过梯度驱动网络输出更具有代表性的语义区域的特征。在这种两方的相互博弈下，网络 N_θ 将为配准头提供更具代表性的特征，以获得更好的对应关系的预测能力，而配准头 G_ξ 也将用有更强大的能力来学习图像之间的相似性，从而反向促进网络表征。

3.2. Z 形匹配模块用于同时学习全局和局部表征

如图4b）所示，Z 形匹配模块 G_ξ 是一种新型的配准学习头，它协同学习仿射和弹性形变配准，同时学习全局和局部的表征能力。它有两个子头，包括用于学习全局相似性的仿射配准头 G_ξ^g 和用于学习局部相似性的弹性配准头 G_ξ^l。

3.2.1 仿射配准头用于学习全局视觉相似性这部分首先将两张图像的编码器部分输出的全局特征（f_A^g、f_B^g）连接起来，然后输入仿射配准头 G_ξ^g 中以预测仿射变换矩阵数值。在我们的研究中，这些数值包括 3 个与旋转相关的数值（θ_x、θ_y、θ_z），3 个与平移相关的数值（t_x、t_y、t_z），3 个与缩放相关的数值（s_x、s_y、s_z）以及 6 个与剪切相关的数值（sh_{xy}、sh_{xz}、sh_{yx}、sh_{yz}、sh_{zx}、sh_{zy}）。这些数值一起计算获得仿射变换矩阵 $\psi_{AB}^g \triangleq G_\xi^g(f_A^g, f_B^g)$，该矩阵表示在空间坐标系中对 x_A 进行全局的变换，从而与 x_B 在全局上对齐。因此，该仿射配准头能够促进模型感知全局对应关系，训练编码器提取相同全局语义区域间一致且具有代表性的特征。

3.2.2 弹性配准头用于学习局部视觉相似性这部分首先利用仿射变换矩阵 ψ_{AB}^g 将网络 N_θ 解码器部分输出的局部特征 f_A^l 在全局上与另一张图像的局部特征 f_B^l 对齐，接着将其与 f_B^l 在特征维度连接起来输入到弹性配准头 G_ξ^l 中以预测用于局部对齐的弹性形变场 $\psi_{AB}^l \triangleq G_\xi^l(\psi_{AB}^g(f_A^l), f_B^l)$。从而将图像 A 形变以实现将其体素对齐到图像 B。在整个过程中，弹性配准头为了学习精准的体素级配准，优化器将训练整个网络 N_θ 对相同局部语义区域提取一致且有代表性的特征。最终，仿射变换矩阵 ψ_{AB}^g 和弹性形变场 ψ_{AB}^l 通过 $\odot$ 操作融合，得到位移向量场 $\psi_{AB} \triangleq \psi_{AB}^l \odot \psi_{AB}^g$ 对每个像素进行位移。（更多的细节在我们的补充材料中。）

Therefore, the correspondence of the visual semantic regions between two images is predicted in ψ_{AB}, and the learning of correspondence will constrain the network $\mathcal{N}_\theta$ to extract more consistent and representative features for same visual semantics, thus driving the head $\mathcal{G}_\xi$ to have more powerful ability to discover their correspondence.

3.3. 基础代理任务用于网络预热

模型对语义区域的基本表征能力对于 GVSL 很重要，因此我们将自修复 [43] 作为框架中的基本预训练任务，与 GVSL 一同进行。这是因为在学习配准的过程中，对应关系的学习 G_ξ 依赖于网络 N_θ 对于两个图像的特征提取能力。然而，如果初始的网络 N_θ 对于特征的表征能力较弱，那么将无法为配准头提供学习发现语义区域间对应关系的特征基础，最终导致发掘可靠的优化目标来训练网络对齐相同的语义区域，阻碍了图像间相似性的学习。因此，我们在框架中融合了一个自修复借口任务 [43] 作为基础学习任务来预热网络。

如图4 c）所示，自修复任务从我们的变换集中采样一个变换操作（$t' \sim T$）来随机变换图像的外观（x_A）以生成变换后的图像（$x_A^{t'} \triangleq t'(x_A)$），然后将其输入到网络 N_θ 中，并从解码器中输出图像的局部特征 $f_A^{l'}$。（为了节省计算资源，在我们的实验中，GVSL 共享此操作，即 $N_\theta(x_A^{t'}) = N_\theta(x_A^t)$。）然后，将特征 $f_A^{l'}$ 输入到一个自修复头 R_ι 中，生成修复后的图像（x'_A），并与原始图像 x_A 一起计算均方误差作为损失函数 $L_{\theta,\iota}^{MSE} = |R_\iota\ (N_\theta(t'(x_A))) - x_A|^2$ [43]，以学习从上下文中修复语义区域原本的特征。因此，网络将在训练过程中学习基本的语义表征来进行预热，即 $\theta, \iota \leftarrow optimizer(\theta, \iota, \nabla_{\theta,\iota} L_{\theta,\iota}^{MSE}, \eta)$，从而避免仅学习 GVSL 时产生无效优化的风险。

4. 实验设置与结果分析

4.1. 实验方案

1）数据集：本实验使用了五个数据集，包括一个私有的预训练数据集和四个公开的下游任务数据集。**预训练数据集**：我们从南京军总医院收集了 302 位患者的无标注心脏 CTA 图像作为自监督预训练数据集。这些图像通过 SOMATOM Definition Flash CT 扫描仪上采集获得，在扫描过程中病人被注射造影剂以获得清晰的心脏内部结构。图像的 x 轴和 y 轴分辨率在 0.25 至 0.57mm/体素之间，大小为 512 个体素，z 轴层厚在 0.75 至 3mm/体素之间，大小在 128 至 994 个体素之间。**下游任务数据集**：我们使用四个公共数据集（MM-WHS-CT [44]、ASOCA [7]、CANDI [20]、STOIC [29]）来验重我们框架的优越性。根据数据种类，我们将其分为场景内迁移和场景间迁移两种不同的研究类型。**场景内迁移**评估旨在利用心脏 CTA 图像上的心脏腔室分割（SHC）[44]、心脏 CTA 图像上的冠状动脉分割（SAC）[7] 以及胸部 CT 图像上的新冠肺炎诊断（CCC）[29] 来评估在与预训练数据集（心脏或胸部 CT）相同/相近图像场景上的迁移能力。**场景间迁移**评估旨在利用脑 MR 图像上的脑组织分割（SBM）[20] 来评估在与源数据集（脑部 MR）不同图像场景上的迁移能力。(更多细节请参考补充材料)

2）对比设置：在这项研究中，我们对比了八个相关的研究工作来验证我们框架的前沿性与优越性，其中包括基于生成的方法 [21, 28, 38, 43]，基于对比的方法 [4, 5, 8]，以及基于聚类的方法 [3] 从而全面分析和展示我们方法在相关研究中的优越性。在所有的实验和对比模型中，我们统一使用 3D U-Net [6] 作为所有方法的骨干网络（全局预测方法使用主干网络的编码器部分）进行公平的比较。我们同时进行了线性评估和微调评估两种评估方式来验证预训练方法的表征能力和迁移能力。

3）评价指标：我们使用平均 Dice 系数（DSC）来评估分割任务性能，使用曲线下面积（AUC）来评估分类任务性能 [35]。

4）实施细节：为了进行公正的比较，所有网络都采用 Adam 优化器来进行训练，学习率设置为 10^{-4}。所有的对比试验都使用 Pytorch [27] 实践，并在一张显存为 24GB 的 NVIDIA Ge-Force RTX 3090 GPU 上进行训练。在预训练任务训练过程中，所有方法都执行了 2×10^5 次迭代，在下游任务训练过程中，所有方法将进行 4×10^4 训练次迭代，并且每执行 200 次迭代，模型都将在验证集上测试性能并保存最佳模型。

4.2. 对比分析

如表1所示，我们的线性 (a) 和微调 (b) 评估显示了我们强大的表征能力和优秀的迁移能力。

4.2.1 同时适用大小结构的场景内迁移能力在场景内迁移学习任务中，GVSL 强大的性能展现出在医学图像大数据量但少标签场景中的巨大应用潜力。在与预训练数据集的相同/近似的图像场景中，GVSL 同时在大结构和小结构分割任务中均取得了最佳性能。**1）**在分割大型心脏结构的 SHC 任务中，GVSL 在线性评估中取得了最高的 DSC（68.4%），比排名第二高的方法 [5] 高出 11.9%。这是因为 GVSL 学习图像间相似性，能

表 1. 线性评估（a）和微调评估（b）展示了 GVSL 强大的表征和迁移能力。表格中加粗的数值代表是该列中的最高值，红色或蓝色的值是指标与随机初始化相比增长或减少的数值。

预训练方法	a）线性评估展现出 GVSL 强大的表征能力				b）微调评估展现出 GVSL 强大的迁移能力			
	$SHC_{DSC\%}$	$SAC_{DSC\%}$	$CCC_{AUC\%}$	$SBM_{DSC\%}$	$SHC_{DSC\%}$	$SAC_{DSC\%}$	$CCC_{AUC\%}$	$SBM_{DSC\%}$
	场景内迁移			场景间迁移	场景内迁移			场景间迁移
随机初始化	21.9	10.0	52.7	56.4	87.8	80.4	74.4	89.7
去噪 [38]	31.4$_{(+9.5)}$	9.3$_{(-0.7)}$	57.9$_{(+5.2)}$	28.3$_{(-28.1)}$	90.3$_{(+2.5)}$	80.5 $_{(+0.1)}$	75.6$_{(+1.2)}$	89.7
抠图 [28]	32.3$_{(+10.4)}$	5.9$_{(-4.1)}$	57.1$_{(+4.4)}$	25.0$_{(-31.4)}$	90.4$_{(+2.6)}$	80.3 $_{(-0.1)}$	79.9$_{(+5.5)}$	89.9$_{(+0.2)}$
Models Genesis [43]	47.4$_{(+25.5)}$	22.5$_{(+12.5)}$	60.4$_{(+7.7)}$	44.9$_{(-11.5)}$	90.3$_{(+2.5)}$	79.9 $_{(-0.5)}$	80.7$_{(+6.3)}$	89.4$_{(-0.3)}$
旋转 [21]	56.1$_{(+34.2)}$	21.9$_{(+11.9)}$	**62.1**$_{(+9.4)}$	54.1$_{(-2.3)}$	90.6$_{(+2.8)}$	81.1$_{(+0.7)}$	77.1$_{(+2.7)}$	89.6$_{(-0.1)}$
深度聚类 [3]	55.9$_{(+34.0)}$	4.4$_{(-5.6)}$	57.9$_{(+5.2)}$	67.5$_{(+11.1)}$	85.4$_{(-2.4)}$	80.5$_{(+0.1)}$	59.9$_{(-14.5)}$	89.1$_{(-0.6)}$
SimSiam [5]	56.5$_{(+34.6)}$	9.7$_{(-0.3)}$	61.0$_{(+8.3)}$	66.2$_{(+9.8)}$	87.5$_{(-0.3)}$	80.1 $_{(-0.3)}$	73.6$_{(-0.8)}$	89.8$_{(+0.1)}$
BYOL [8]	46.9$_{(+25.0)}$	8.6$_{(-1.4)}$	53.7$_{(+1.0)}$	52.7$_{(-3.7)}$	88.6$_{(+0.8)}$	80.7 $_{(+0.3)}$	76.5$_{(+2.1)}$	89.5$_{(-0.2)}$
SimCLR [4]	48.7$_{(+26.8)}$	15.5$_{(+5.5)}$	61.3$_{(+8.6)}$	58.7$_{(+2.3)}$	86.9 $_{(-0.9)}$	79.9$_{(-0.5)}$	74.3$_{(-0.1)}$	89.3$_{(-0.4)}$
去除“Z 形匹配模块”	49.1$_{(+27.2)}$	21.1$_{(+11.1)}$	55.8$_{(+3.4)}$	45.1$_{(-11.3)}$	88.3$_{(+0.5)}$	81.2$_{(+0.8)}$	81.3$_{(+6.9)}$	89.7
去除“自修复”	45.3$_{(+23.4)}$	0.0$_{(-10.0)}$	58.8$_{(+6.4)}$	48.5$_{(-7.9)}$	87.0$_{(-0.8)}$	79.5 $_{(-0.9)}$	76.6$_{(+2.2)}$	89.0$_{(-0.7)}$
去除“仿射配准模块”	57.7$_{(+35.8)}$	17.9$_{(+7.9)}$	57.6$_{(+4.9)}$	53.4$_{(-3.0)}$	89.4$_{(+1.6)}$	**82.3**$_{(+1.9)}$	79.8$_{(+5.4)}$	89.8$_{(+0.1)}$
完整的 GVSL	**68.4**$_{(+46.5)}$	**28.7**$_{(+18.7)}$	60.8$_{(+8.1)}$	**79.9**$_{(+23.5)}$	**91.2**$_{(+3.4)}$	81.3$_{(+0.9)}$	**82.2**$_{(+7.8)}$	**90.0**$_{(+0.3)}$

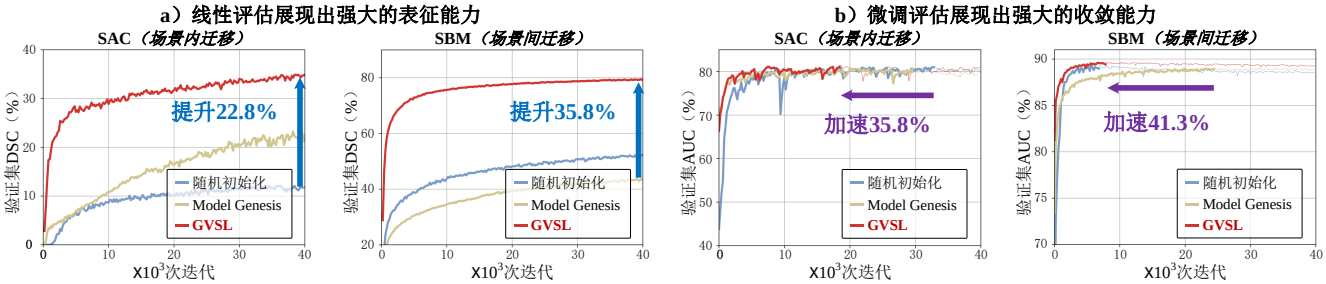


图 6. GVSL 在线性评估中展现出了强大的表征能力，在微调评估中展现出了更快的收敛能力。在 b）中，粗线的部分表示验证集上未达到最佳模型的训练过程。

够促进语义区域学习一致的特征表征。特别是对于具有清晰语义区域的大解剖结构，如 CTA 图像上的心脏内部腔室，配准学习能够带来更加高效的表征学习。**2)** 在分割细小冠脉的 SAC 任务中，大量已有的预训练方法 [4,5,8,28] 都会导致模型忽略这些细小结构的特征，导致模型性能甚至低于随机初始化的模型。特别是为全局预测所设计的方法 [4,5,8]，这些方法更关注全局大感受野下特征的学习，更容易丢失对细小结构的表征。我们 GVSL 中的弹性形变配准学习和自修复学习训练网络学习精细的细小视觉语义特征表征，取得了线性评估中的最高 DSC（28.7%）。有趣的是，在我们的 Z 形匹配模块中移除仿射配准头后，GVSL 进一步在 SAC 任务中取得了最佳的微调评估性能（82.3%），进一步证明了为细小物体学习密集表征的重要性。

4.2.2 强大的场景间迁移能力在场景间迁移学习任务中，GVSL 进一步展现出强大的性能，说明了我们的作为预训练参数初始化网络的有效性。SBM 任务使用了与预训练图像不同模态和图像上下文（身体范围）的脑 MR 图像，构成了场景间迁移的评估。在线性评估中，大量的对比方法 [8,8,21,43] 由于预训练数据集（心脏 CT）和下游任务数据集（脑 MR）之间巨大的差异而使得性能甚至弱于随机初始化。而我们的 GVSL 在线性评估中取得了最高的 DSC（79.9%），与随机初始化相比提高了 23.5%。这是因为 GVSL 学习图像间相似性，使得预训练网络对语义区域拥有了更加一致的表征，使得特征拥有更好的聚类特性，在该特性下模型的表征将更加容易整体地迁移至目标下游任务。但是值得注意的是，在微调评估中，所有方法的性能都与随

机初始化性能相似甚至更差 [3,4,8,21,43]。这是因为脑 MR 和心脏 CTA 图像之间有着巨大的分布差异（不同的模态和身体范围），使得预训练模型的表征能力与下游任务之间相去甚远，无法获得有贡献的迁移。但是即便如此，GVSL 仍然取得了 0.3% 的提升。

4.2.3 全局和密集预测任务的优越性 GVSL 在全局预测和密集预测任务中都拥有极大的性能优越性，展示出其在不同类型潜在下游任务中的适应性能力。**1)** 在密集预测任务中（SHC，SAC，SBM），由于我们的弹性配准头和自修复头训练模型学习局部密集特征的表征能力，GVSL 在所有三个任务中都取得了最高的 DSC。SimSiam、BYOL、SimCLR 和深度聚类只学习全局表征，因此对于细小结构的表征能力较差，SAC 任务中性能极差。**2)** 在全局预测任务中（CCC），由于医学图像间语义内容的一致性，其在全局上表现出极大的相似性，而任务所关注病变发生在局部区域。因此，GVSL 利用 Z 形匹配模块同时学习全局和局部的视觉相似性，实现对全局和局部特征的强大表征能力，因此在微调评估中获得了最高的 AUC（82.2%），证明了我们方法在全局预测任务的强大迁移能力。尽管 GVSL 在线性评估中比最高方法（旋转）低 1.3%，但仍高于那些仅学习局部密集表征的预训练网络的，包括去噪、抠图和 Model Genesis。

4.3. 消融实验和模型分析

4.3.1 消融实验如表1所示，我们将 GVSL 与仅学习基础预训练任务（自修复任务）、仅用使用 Z 形匹配模块学习配准和基础预训练任务 + 仅实现弹性配准头学习配准这三种方法进行比较。我们可以得到三个观察结论：**1)** 仅学习 Z 形匹配模块时，网路较弱的初始表征能力使得模型的优化非常低效，导致较差的表征能力。特别是在 SAC 任务的线性评估中，由于简单的配准学习导致的最终获得表征的低效性，在这类细小结构分割任务上无法豁达有效的性能。**2)** 当仅学习基础" 自修复 "任务时，模型整体性能较差，尤其是 SAC 任务上，无论是线性评估还是微调评估，该模型性能均比随机初始化更差。**3)** 当删除 Z 形匹配模块中的仿射头时，由于缺乏全局表征学习，尽管该模型在 CCC 任务的线性评估和微调评估的 AUC 由于随机初始化，但是他们均弱于我们最终的 GVSL 框架，分别降低了 3.2% 和 2.4%。由此可见，我们的每个模块在整个框架中都具有其必要性和有效性。

4.3.2 对学习效率的提升如图6所示，我们分析了随机初始化、GVSL 预训练和 Model Genesis 预训练模型在 SAC 和 SBM 任务中的学习性能。在线性评估中，由于 GVSL 对局部图像间相似性的有效学习，相比于随机初始化和 Model Genesis，我们的方法取得了超过 20% 的 DSC 提升。在微调评估中，GVSL 以其极强的表征能力大大提高了训练的收敛速度，获得了超过 30% 的速度提升，表现出其强大的收敛能力和节省计算资源的巨大潜力。尽管在 SBM 任务的微调评估中，随机初始化的收敛速度似乎更快，但它很快陷入了过度拟合，最终性能弱于我们的方法。

4.3.3 基础任务对于 GVSL 的必要性自修复学习为网络提供了基础的视觉特征表征，进从而驱动配准学习过程中对图像间相似性的学习，具有必要性。我们验证了当去除自修复学习后，模型对于配准学习的收敛性。如图8所示，当仅学习配准借口任务时，网络由于较弱的初始表征能力，使得对于配准的学习效率极为低下，因此用于学习配准的相似性损失（NCC 损失，L^{NCC}）无法收敛，导致 Z 配准头无法学习语义区域的对应关系。当添加我们基础的自修复借口任务时，由于网络从自修复学习中获取了对语义区域的基本表征能力，配准学习被有效驱动，NCC 损失开始下降，从而我们的 GVSL 在整个学习过程中发挥作用，模型能够学习到更有效的语义特征聚类效应。如表1所示，加入自修复学习和仅进行 Z 形匹配模块的配准学习相比，加入自修复学习后性能获得了显著的提升。

4.3.4 GVSL 预训练模型的特征聚类效应如图7所示，我们对 SHC 任务的图像使用预训练模型提取局部特征 f^l，利用 t-SNE 压缩特征后绘制散点图以展示了模型对特征的聚类效应，结果显示我们 GVSL 拥有更好的聚类能力，表现出其强大的表征性能。随机初始化方法无法提取具有代表性的特征，无法区分潜在的语义区域，因此它将不同的语义特征混合，表征能力较差。当只学习自修复任务时，模型能够学习到特征的基本表征能力，但整体模型由于无法学习图像间相似性，表征能力仍然受限。因此，其提取的特征 f^l 仍然存在较严重的混合。当仅使用我们的 Z 形匹配模块学习配准任务时，所提取的特征表现出明显更强的聚类效应，说明了学习图像间相似性的重要性。然而，某些语义特征仍然被混合到一起，无法被分开。当我们融合学习配

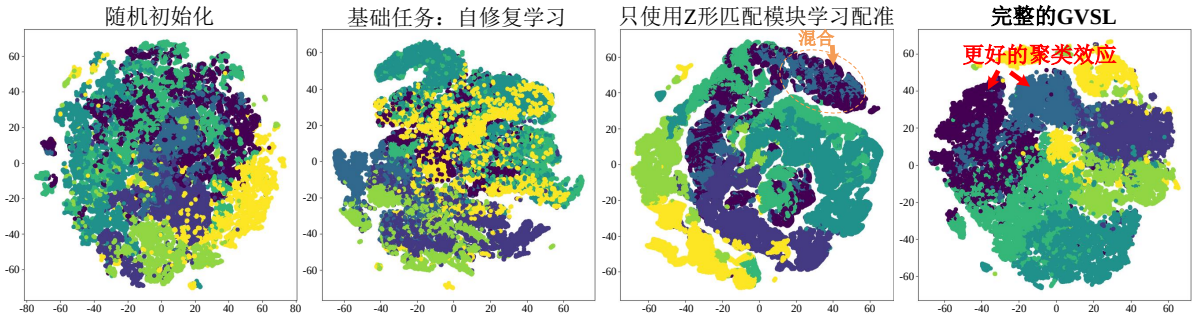


图 7. 通过 t-SNE [36] 将模型提取的特征压缩成 2 维向量后绘制散点图，对预训练网络在 SHC 任务中提取的七个语义区域（AO、RA、RV、LV、PA、LA 和 Myo）的局部特征 f^l 进行的聚类可视化，我们发现 GVSL 表现出了更好的特征聚类效应。

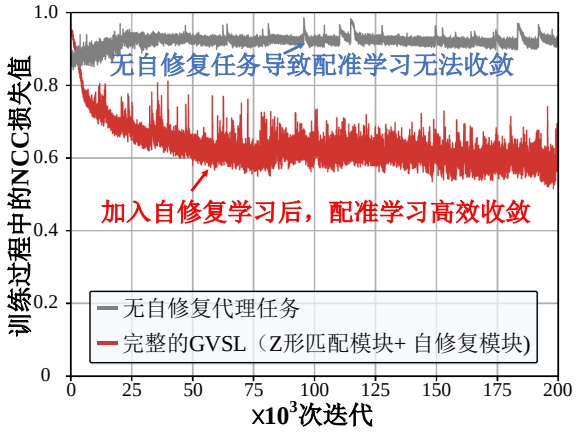


图 8. 我们的 GM 学习中基础任务的必要性。a）当仅训练 GM 任务时，NCC 损失无法收敛，无法学习语义区域的对应关系。b）当添加基础任务（自恢复）时，从语义区域的基本表示进行预热，促进了 GM 的有效优化。

准和学习自修复到一起来训练 GVSL 时，来自基础任务的基本表征能力促进了配准的学习，最终获得了极好的特征聚类效应，不同语义的特征具有较好的区分度。尽管黄色的点图中被划分为了三个部分，但是每个部分都是聚集的，这也说明了该区域内部语义表征的一致性。

5. 总结和讨论

在本文中，我们提出了三维医学图像自监督预训练中的图像间相似性学习，并提出了用于表征图像间相似性的几何视觉相似性学习 (GVSL)，为下游特定应用任务中的迁移学习提供了强大的初始表征能力。虽然其学习相同语义的一致表示的能力在 3D 图像 (CT、MR) 的场景内、场景间和全局-局部迁移学习任务中具有强大的性能，但未来重要的工作是将图像间相似性的学习扩展到一些没有拓扑不变性的图像，类似于病理全切片成像。我们相信，GVSL 将促进医学图像分析中自监督学习的研究，并且作为下游任务的迁移学习的主要参数来源。

对潜在影响的讨论所提出的方法在医学成像分析中展示了有效且合理的潜力，展示出了巨大的潜在影响力。特别是对于被广泛使用的 3D 医学图像，其三维结构的空间完整性避免了二维图像的空间投影 (如 X 射线图像) 和空间遮挡 (如自然图像) 存在的信息畸变和缺失的问题。人体间的一致性也使得这些 3D 图像中的内容具有拓扑不变性。因此，这些图像之间的几何关系能够有效地用于驱动视觉间相似性的度量。我们相信，图像中三维结构的空间完整性和拓扑不变性将进一步启发研究者对自监督预训练进行更多的研究。

对潜在限制的讨论我们的 GVSL 仍然有一些局限性。1）Z 形匹配模块和自恢复学习中的额外计算为模型的训练提出了更大的 GPU 存储空间的需求和更多的计算成本。2）由于源场景和目标场景之间的巨大差异，场景间的迁移对于医学图像预训练模型仍然是一个巨大的挑战。幸运的是，由于 GPU 的发展和医学数据集的扩大 (为构建场景内迁移的情况提供了更多的可能性)，这些限制正在逐步得到解决。

致 谢 本研究得到了国家重点研发计划项目 (2022YFE0116700)、CAAI-华为 MindSpore 开放基金和东南大学研究生院科研基金 (YBPY2139) 的资助。我们感谢东南大学大数据计算中心提供的设施支持。我们还要感谢教育部计算机网络与信息集成重点实验室和江苏省医学信息处理国际联合研究实验

室。

参考文献

[1] Aurélien Bellet, Amaury Habrard, and Marc Sebban. Metric learning. *Synthesis lectures on artificial intelligence and machine learning*, 9(1):1–151, 2015. 4

[2] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. 1

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 1, 2, 3, 6, 7, 8

[4] 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. 2, 3, 6, 7, 8

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, June 2021. 2, 3, 6, 7

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 6

[7] Ramtin Gharleghi, Dona Adikari, Katy Ellenberger, Sze-Yuan Ooi, Chris Ellis, Chung-Ming Chen, Ruochen Gao, Yuting He, Raabid Hussain, Chia-Yen Lee, et al. Automated segmentation of normal and diseased coronary arteries-the asoca challenge. *Computerized Medical Imaging and Graphics*, page 102049, 2022. 6

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Pires, Zhaohan Guo, Mohammad Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Neural Information Processing Systems*, 2020. 2, 6, 7, 8

[9] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–147. Springer, 2020. 3

[10] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE transactions on medical imaging*, 2021. 3

[11] Joon Hee Han and Jong Seung Park. Contour matching using epipolar geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):358–370, 2000. 2

[12] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *Proceedings of the IEEE international conference on computer vision*, pages 1831–1840, 2017. 2, 3

[13] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1):1–18, 2020. 2, 3

[14] 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 (CVPR)*, June 2020. 3

[15] Yuting He, Rongjun Ge, Xiaoming Qi, Yang Chen, Jiasong Wu, Jean-Louis Coatrieux, Guanyu Yang, and Shuo Li. Learning better registration to learn better few-shot medical image segmentation: Authenticity, diversity, and robustness. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2, 3

[16] Yuting He, Tiantian Li, Rongjun Ge, Jian Yang, Youyong Kong, Jian Zhu, Huazhong Shu, Guanyu Yang, and Shuo Li. Few-shot learning for deformable medical image registration with perception-correspondence decoupling and reverse teaching. *IEEE Journal of Biomedical and Health Informatics*, 2021. 2, 3

[17] Yuting He, Tiantian Li, Guanyu Yang, Youyong Kong, Yang Chen, Huazhong Shu, Jean-Louis Coatrieux, Jean-Louis Dillenseger, and Shuo Li. Deep complementary joint model for complex scene registration and few-shot segmentation on medical images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 770–786. Springer, 2020. 3

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 3, 4

[19] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 3

[20] David N Kennedy, Christian Haselgrove, Steven M Hodge, Pallavi S Rane, Nikos Makris, and Jean A Frazier. Candishare: A resource for pediatric neuroimaging data. *Neuroinformatics*, 10(3):319, 2012. 6

[21] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 3, 6, 7, 8

[22] Xiaoni Li, Yu Zhou, Yifei Zhang, Aoting Zhang, Wei Wang, Ning Jiang, Haiying Wu, and Weiping Wang. Dense semantic contrast for self-supervised visual representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1368–1376, 2021. 1, 2, 3

[23] 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. 2, 3

[24] Timo Milbich. *Visual Similarity and Representation Learning*. PhD thesis, 2021. 1, 3

[25] Michael I. Miller and Laurent Younes. Group actions, homeomorphisms, and matching: A general framework. *International Journal of Computer Vision*, 41(1):61–84, 2001. 2

[26] Frank H Netter. *Atlas of human anatomy, Professional Edition E-Book: including NetterReference. com Access with full downloadable image Bank*. Elsevier health sciences, 2014. 1, 2

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 6, 7

[29] Marie-Pierre Revel, Samia Boussouar, Constance de Margerie-Mellon, Inès Saab, Thibaut Lapotre, Dominique Mompoint, Guillaume Chassagnon, Audrey Milon, Mathieu Lederlin, Souhail Bennani, et al. Study of thoracic ct in covid-19: The stoic project. *Radiology*, 301(1):E361–E370, 2021. 6

[30] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017. 2, 3

[31] Karsten Roth, Timo Milbich, and Bjorn Ommer. Pads: Policy-adapted sampling for visual similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6568–6577, 2020. 1

[32] Walid Saad, Zhu Han, Mérouane Debbah, Are Hjorungnes, and Tamer Basar. Coalitional game theory for communication networks. *Ieee signal processing magazine*, 26(5):77–97, 2009. 5

[33] Chaman L Sabharwal and Jennifer L Leopold. A completeness of metrics for topological relations in 3d qualitative spatial reasoning. *Polibits*, 52:5–15, 2015. 1

[34] Jiacheng Shi, Yuting He, Youyong Kong, Jean-Louis Coatrieux, Huazhong Shu, Guanyu Yang, and Shuo Li. Xmorpher: Full transformer for deformable medical image registration via cross attention. *arXiv preprint arXiv:2206.07349*, 2022. 3

[35] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015. 6

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

[37] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc V Gool. Revisiting contrastive methods for unsupervised learning of visual representations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16238–16250. Curran Associates, Inc., 2021. 3

[38] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 2, 3, 6, 7

[39] 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. 2

[40] Eric W Weisstein. Affine transformation. *https://mathworld. wolfram. com/*, 2004. 3

[41] 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*, 41(10):2658–2669, 2022. 3

[42] Borui Zhang, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Attributable visual similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7532–7541, 2022. 1

[43] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pages 384–393. Springer, 2019. 1, 2, 3, 6, 7, 8

[44] Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58:101537, 2019. 6