

# 目录

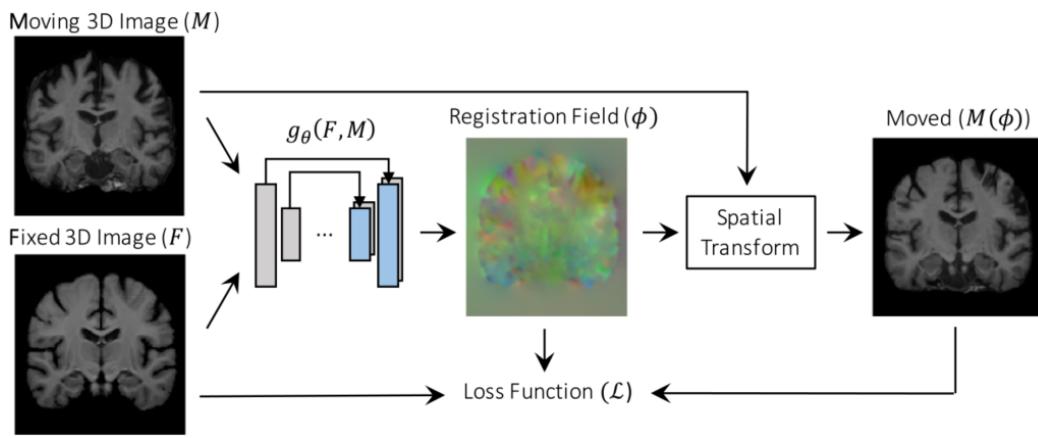
---

## 目录

1. VoxelMorph: A Learning Framework for Deformable Medical Image Registration (2019 CVPR VoxelMorph)
2. A coarse-to-fine deformable transformation framework for unsupervised multi-contrast MR image registration with dual consistency constraint (2020 TMI)
3. A deep learning framework for unsupervised affine and deformable image registration (2019 Medical Image Analysis DLIR)
4. Attention for Image Registration (AiR): an unsupervised Transformer approach (2021 arxiv AiR)
5. BIRNet: Brain image registration using dual-supervised fully convolutional networks (2018 MedIA BIRNet)
6. End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network (2017 DLMIA&ML-CDS&MICCAI DIRNet)
7. Inverse-Consistent Deep Networks for Unsupervised Deformable Image Registration (2018 arxiv ICNet)
8. Learning a Deformable Registration Pyramid (2021 MICCAI)
9. Unsupervised 3D End-to-End Medical Image Registration with Volume Tweening Network (2019 JBHI VTN)
10. Recursive Cascaded Networks for Unsupervised Medical Image Registration (2019 ICCV)
11. Affine Medical Image Registration with Coarse-to-Fine Vision Transformer (2022 CVPR C2FViT)
12. Pyramid Vision Transformer A Versatile Backbone for Dense Prediction without Convolutions (2021 ICCV PVT)
13. TransMorph: Transformer for unsupervised medical image registration (2021 MedIA TransMorph)
14. CycleMorph: Cycle consistent unsupervised deformable image registration (2020 MedIA CycleMorph)
15. ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration (2021 arxiv ViT-V-Net)
16. Dual-stream pyramid registration network (2020 MICCAI/Media Dual-PRNet)
17. XMorpher: Full Transformer for Deformable Medical Image Registration via Cross Attention (2022 MICCAI XMorpher)
18. Learning Dual Transformer Network for Diffeomorphic Registration (2021 MICCAI)
19. DeepAtlas: Joint Semi-supervised Learning of Image Registration and Segmentation (2019 MICCAI)
20. A Cross-Stitch Architecture for Joint Registration and Segmentation in Adaptive Radiotherapy (2020 PMLR)
21. A Hybrid Deep Learning Framework for Integrated Segmentation and Registration: Evaluation on Longitudinal White Matter Tract Changes (2019 MICCAI)
22. A segmentation-informed deep learning framework to register dynamic two-dimensional magnetic resonance images of the vocal tract during speech (2022 BSPC)
23. Deep Complementary Joint Model for Complex Scene Registration and Few-Shot Segmentation on Medical Images (2020 ECCV)
24. Deep Learning-Based Concurrent Brain Registration and Tumor Segmentation (2019 Frontiers in Computational Neuroscience)
25. Image-and-Spatial Transformer Networks for Structure-Guided Image Registration (2019 MICCAI)

## 1. VoxelMorph: A Learning Framework for Deformable Medical Image Registration (2019 CVPR VoxelMorph)

### 1. 方法



- 使用U-net得到flow field,然后使用spatial transformer得到配准图像;
- ncc loss和梯度平滑loss;
- 有分割图像作为辅助信息,计算分割图像配准前后的dice作为总loss的一部分。

## 2. 代码

- 生成flow field的网络的最后一层使用0均值, 1e-5方差的正态分布初始化参数。

```

1 | self.flow.weight = nn.Parameter(Normal(0, 1e-
5).sample(self.flow.weight.shape))
2 | self.flow.bias =
nn.Parameter(torch.zeros(self.flow.bias.shape))

```

- flow field和标准网格相加然后再归一化到 (-0.1,0.1) 。

```

1 | def forward(self, src, flow):
2 |     # new locations
3 |     new_locs = self.grid + flow
4 |     shape = flow.shape[2:]
5 |
6 |     # need to normalize grid values to [-1, 1] for resampler
7 |     for i in range(len(shape)):
8 |         new_locs[:, i, ...] = 2 * (new_locs[:, i, ...] /
(shape[i] - 1) - 0.5)
9 |
10 |     # move channels dim to last position
11 |     # also not sure why, but the channels need to be reversed
12 |     if len(shape) == 2:
13 |         new_locs = new_locs.permute(0, 2, 3, 1)
14 |         new_locs = new_locs[..., [1, 0]]
15 |     elif len(shape) == 3:
16 |         new_locs = new_locs.permute(0, 2, 3, 4, 1)
17 |         new_locs = new_locs[..., [2, 1, 0]]
18 |
19 |     return nnf.grid_sample(src, new_locs, align_corners=True,
mode=self.mode)

```

- 医学图像处理包 nibabel

## 3. 问题:

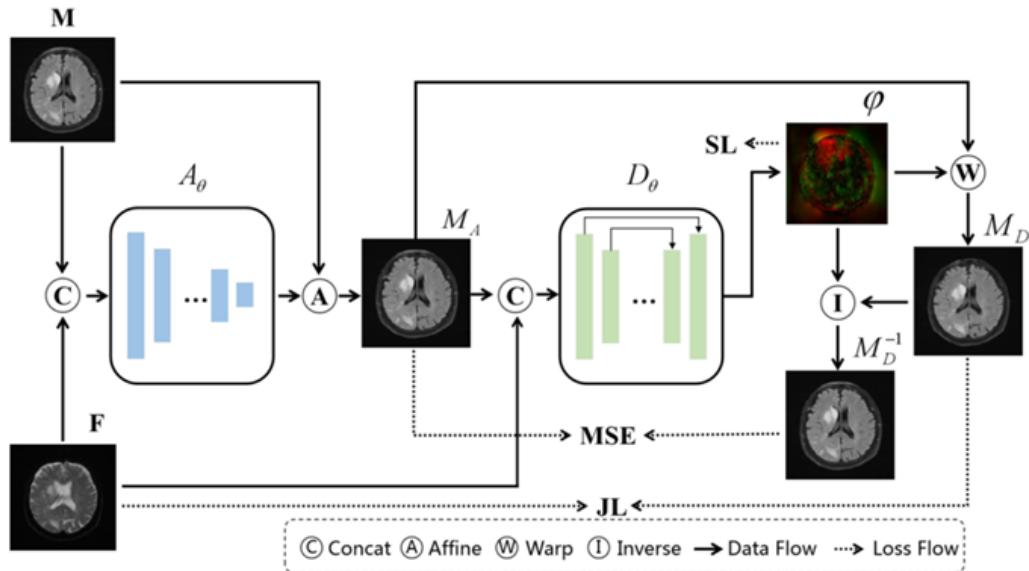
代码VoxelMorph-torch/Model/model.py/forward()中new\_locs = new\_locs[...,[2, 1, 0]],为什么要交换x,y,z的位置;

## 2. A coarse-to-fine deformable transformation framework for unsupervised multi-contrast MR image registration with dual consistency constraint (2020 TMI)

### 1. 动机

- 先叙述了多对比图像和多对比图像分析的意义，然后指出了传统的多对比分析方法的缺点，最后直接说提出了这个框架。

### 2. 方法



- 先使用一个网络输出仿射变换参数进行仿射变换，然后使用Unet结构输出flow field，然后进行配准操作。

- 创新点：逆变换场，和两个损失函数。

- 提出JL loss，即除了similarity (这里是MI) term之外对于图像background 的配准结果计算MSE，抑制warped image中处于fixed image background 区域中出现object (作者基于磁共振背景的灰度信号接近于0提出了背景抑制损失函数，该函数将f中灰度值小于某个值 $\gamma$  (由数据集据统计得到) 的部分特别额外做一个MSE运算 (其实也就是相当于让配准后图像对应位置也应该0) )。

Combing the MI loss function and the prior knowledge-based background suppressing loss function, we obtain the first loss function, which is called a prior knowledge-based joint loss function ( $JL(F, \xi_\theta(F, M), \lambda)$ ) as shown in Eq.6:

$$JL(F, \xi_\theta(F, M), \alpha, \beta) = \sum_{f, m} (\alpha MI(f, \xi_\theta(f, m)) + \beta \sum_i \begin{cases} MSE(f_i, \xi_\theta(f, m)_i), & \text{if } f_i < \gamma \\ 0, & \text{otherwise} \end{cases}) \quad (6)$$

where  $i \in N$  represents the pixels in images,  $\gamma$  is a threshold obtained from the data set to determine whether the pixel is background or not,  $\alpha$  and  $\beta$  are adjust factors to balance the two losses. JL can not only constrain the global image alignment by maximizing MI, but also penalize the incorrect predictions in defined regions. This makes the predictions more in line with the nature of medical images.

- 双重一致性损失：变形配准后的图像经过逆变换后和配准前的图像进行MSE或NCC的损失。
- 总loss等于上述两项loss加上梯度平滑loss（下图）：

$$\mathcal{L}_{smooth}(\phi) = \sum_{\mathbf{p} \in \Omega} \|\nabla \mathbf{u}(\mathbf{p})\|^2, \quad (7)$$

and approximate spatial gradients using differences between neighboring voxels. Specifically, for  $\nabla \mathbf{u}(\mathbf{p}) = \left( \frac{\partial \mathbf{u}(\mathbf{p})}{\partial x}, \frac{\partial \mathbf{u}(\mathbf{p})}{\partial y}, \frac{\partial \mathbf{u}(\mathbf{p})}{\partial z} \right)$ , we approximate  $\frac{\partial \mathbf{u}(\mathbf{p})}{\partial x} \approx \mathbf{u}((p_x + 1, p_y, p_z)) - \mathbf{u}((p_x, p_y, p_z))$ , and use similar approximations for  $\frac{\partial \mathbf{u}(\mathbf{p})}{\partial y}$  and  $\frac{\partial \mathbf{u}(\mathbf{p})}{\partial z}$ .

### 3. 代码

有，没看完。对于形变场中的一个点（位移向量），如果它的雅克比行列式小于0，则表示该点发生了折叠（folding），代码如下：

```

1 def jacobian_determinant(disp):
2     """
3         jacobian determinant of a displacement field.
4         NB: to compute the spatial gradients, we use np.gradient.
5         Parameters:
6             disp: 2D or 3D displacement field of size [*vol_shape,
7                 nb_dims],
8                     where vol_shape is of len nb_dims
9         Returns:
10            jacobian determinant
11            """
12
13     # check inputs
14     volshape = disp.shape[:-1]
15     nb_dims = len(volshape)
16     assert len(volshape) in (2, 3), 'flow has to be 2D or 3D'
17
18     # compute grid
19     grid_1st = nd.volsize2ndgrid(volshape)
20     grid = np.stack(grid_1st, len(volshape))
21
22     # compute gradients
23     J = np.gradient(disp + grid)
24
25     # 3D glow
26     if nb_dims == 3:
27         dx = J[0]
28         dy = J[1]
29         dz = J[2]
30
31         # compute jacobian components
32         Jdet0 = dx[..., 0] * (dy[..., 1] * dz[..., 2] - dy[..., 2] *
33         dz[..., 1])
34         Jdet1 = dx[..., 1] * (dy[..., 0] * dz[..., 2] - dy[..., 2] *
35         dz[..., 0])
36         Jdet2 = dx[..., 2] * (dy[..., 0] * dz[..., 1] - dy[..., 1] *
37         dz[..., 0])
38
39         return Jdet0 - Jdet1 + Jdet2
40
41     else: # must be 2

```

```

38         dfdx = J[0]
39         dfdy = J[1]
40     return dfdx[..., 0] * dfdy[..., 1] - dfdy[..., 0] * dfdx[...,
1]

```

#### 4. 问题

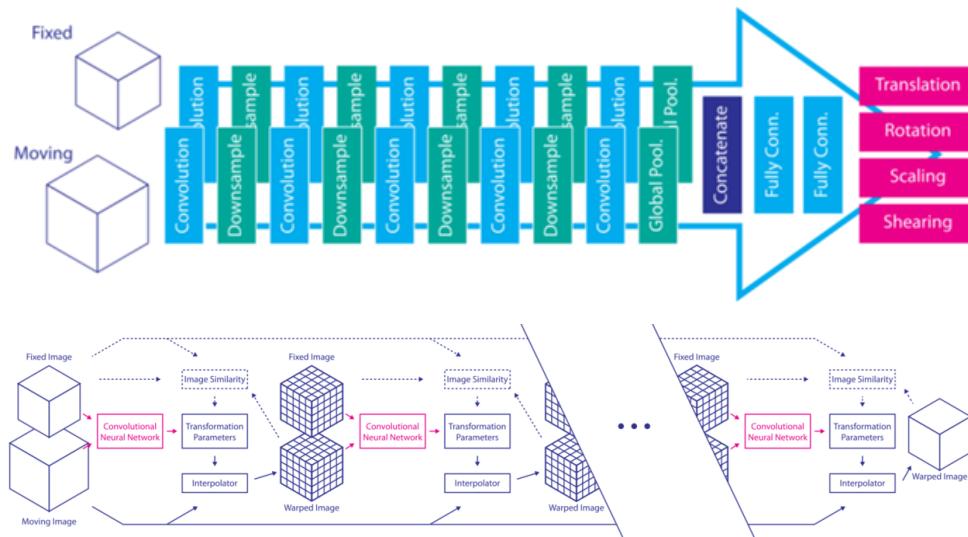
逆形变场的理解。

### 3. A deep learning framework for unsupervised affine and deformable image registration (2019 Medical Image Analysis DLIR)

#### 1. 动机

作者列举出了一些基于深度学习的方法，指出了这些方法虽然显示了准确的配准性能，但这些方法都是有监督的，即它们依赖于示例配准进行训练或需要手动分割。然后，又叙述了无监督DL方法已经用于光流估计领域。最后作者直接说明了提出了一种深度学习图像配准(DLIR)框架：一种无监督技术，用于训练CNN进行医学图像配准任务。

#### 2. 方法



- 仿射变换网络输出12个仿射变换参数：三个平移，三个旋转，三个缩放和三个剪切参数；
- 按顺序训练，每个阶段都为其特定的注册任务进行训练，同时保持前几个阶段的权重固定；
- ncc loss和bending energy penalty loss:

$$L = L_{NCC} + \alpha P, \quad (1)$$

where  $L_{NCC}$  is the negative normalized cross correlation, and  $P$  the bending energy penalty with  $\alpha = 0$  for affine registration, and  $\alpha$  empirically determined to be 0.05 for all deformable image registration experiments. The bending energy penalty is defined as follows:

$$\begin{aligned} P = \frac{1}{V} \int_0^X \int_0^Y \int_0^Z & \left[ \left( \frac{\partial^2 \mathbf{T}}{\partial x^2} \right)^2 + \left( \frac{\partial^2 \mathbf{T}}{\partial y^2} \right)^2 + \left( \frac{\partial^2 \mathbf{T}}{\partial z^2} \right)^2 \right. \\ & \left. + 2 \left( \frac{\partial^2 \mathbf{T}}{\partial xy} \right)^2 + 2 \left( \frac{\partial^2 \mathbf{T}}{\partial xz} \right)^2 + 2 \left( \frac{\partial^2 \mathbf{T}}{\partial yz} \right)^2 \right] dx dy dz, \end{aligned}$$

where  $V$  is the volume of the image domain, and  $\mathbf{T}$  the local transformation. Adding this term during registration minimizes the second order derivatives of local transformations of a DVF, thereby resulting in locally affine transformations, thus enforcing global smoothness (Staring et al., 2007):

### 3. 代码

无

### 4. 问题

~~不能理解bending energy penalty loss。~~

bending energy penalty loss:

$$\begin{aligned} \mathcal{R}_{bending}(\phi) = \sum_{\mathbf{p} \in \Omega} \|\nabla^2 \mathbf{u}(\mathbf{p})\|^2 = \sum_{\mathbf{p} \in \Omega} & \left[ \left( \frac{\partial^2 \mathbf{u}(\mathbf{p})}{\partial x^2} \right)^2 + \left( \frac{\partial^2 \mathbf{u}(\mathbf{p})}{\partial y^2} \right)^2 + \right. \\ & \left. \left( \frac{\partial^2 \mathbf{u}(\mathbf{p})}{\partial z^2} \right)^2 + 2 \left( \frac{\partial^2 \mathbf{u}(\mathbf{p})}{\partial xz} \right)^2 + 2 \left( \frac{\partial^2 \mathbf{u}(\mathbf{p})}{\partial xy} \right)^2 + 2 \left( \frac{\partial^2 \mathbf{u}(\mathbf{p})}{\partial yz} \right)^2 \right], \end{aligned}$$

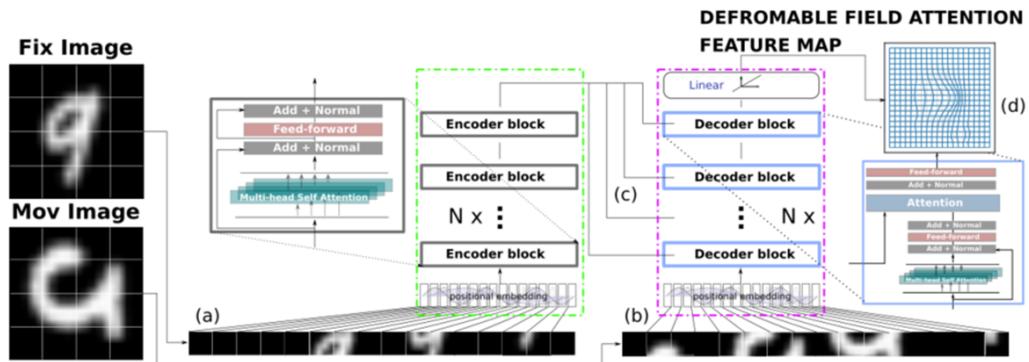
它惩罚急剧弯曲的变形，是作用于位移场的二阶导数。

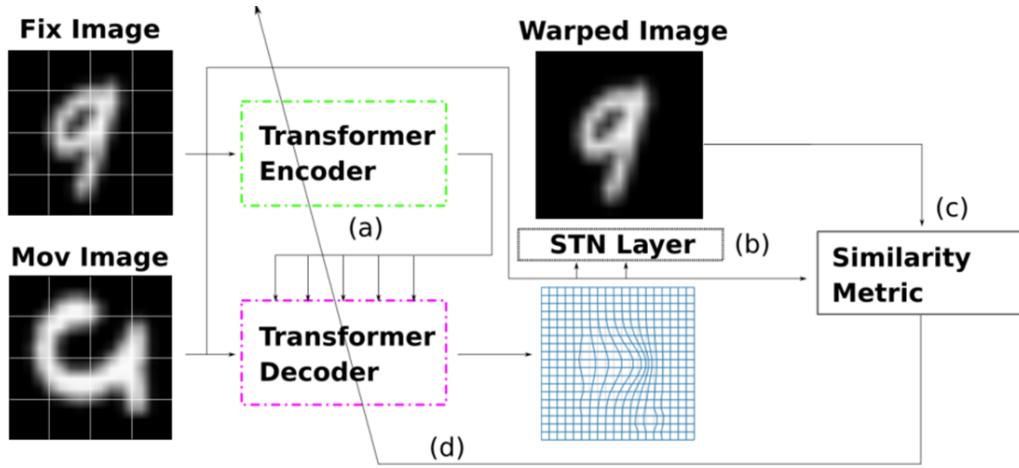
## 4. Attention for Image Registration (AiR): an unsupervised Transformer approach (2021 arxiv AiR)

### 1. 动机

- Transformer在各个领域展示了强大的能力，本文试图将Transformer模型引入图像配准领域，提出的框架是第一个基于Transformer的图像配准方法。

### 2. 方法





- 将配准视为翻译任务，输入移动图像和固定图像，经过一个一个Transformer输出形变场。具体来说，AiR将固定图像分成一些patch序列，输入到Encoder中；将移动图像分成一些patch序列输入到Decoder中，Transformer整体结构不变，最后输出形变场，经过一个STN网络得到配准图像。
- 提出了一种多尺度注意力并行Transformer(MAPT)，它可以从不同的感知尺度学习特征。MAPT由N个Transformer（N个解码器和N个编码器）组成。对于每个变压器，他们采用不同大小的patch作为输入，生成N个不同的注意力特征图 $F_N$ 。然后将N个特征图采样成统一的大小，并按归一化加权比例相加，得到最终的可变形特征图 $F$ 。感觉这个部分论文中没有讲清楚。
- 实验有点少，说服力也不够强。

### 3. 代码

给出了链接，但无法正常访问。

### 4. 问题

无

## 5. BIRNet: Brain image registration using dual-supervised fully convolutional networks (2018 MedIA BIRNet)

### 1. 动机

- 与基于深度学习的配准方法相比，文章旨在解决缺少理想groundtrue形变场（有形变场，但不理想）的问题，进而进一步提高配准精度。用其他方法获得的形变场辅助配准的方法。groundtrue形变场用于快速粗配准，图像相似性损失用于细配准。

### 2. 方法

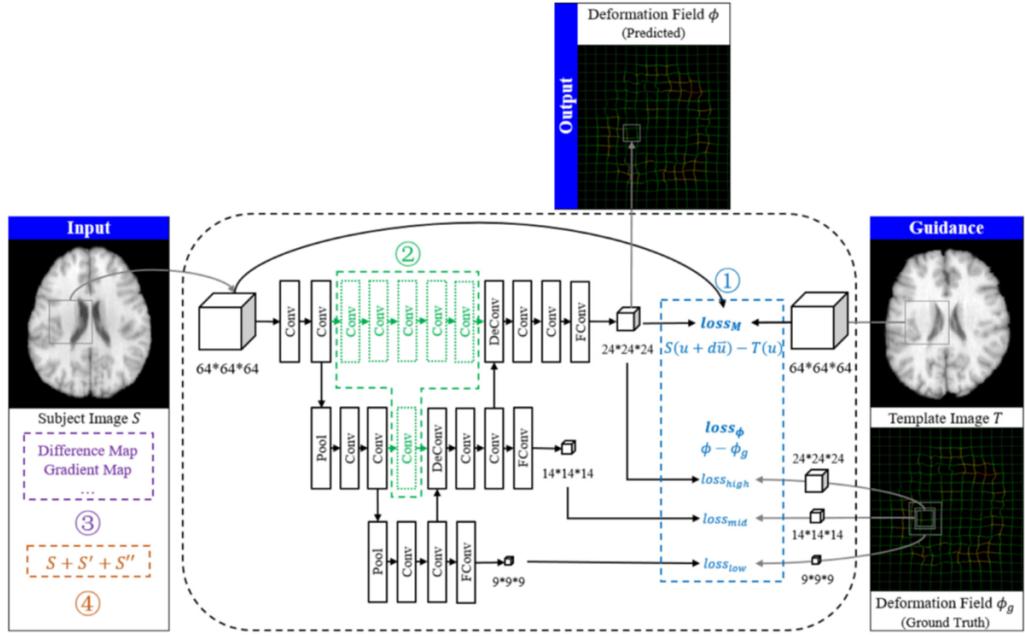


Fig. 1. Overview of our proposed method.

- motivation:
- point:

1. Hierarchical dual-supervision 双重监督策略: 预测变形场与现有groudtrue 变形场之间的差异  $loss_{\phi}$ , 从groudtrue形变场中抽出 $24^3, 14^3, 9^3$ 三种分辨率的patch (方法如下图), 与 U-net对应各层输出的形变场计算loss, 将它们相加得到  $loss_{\phi}$ ; 配准图像与固定图像的差异  $loss_M^*$ 。

$$\phi_g^{high}(i, j, k) = \phi_g(i + 20, j + 20, k + 20) \\ i, j, k \in [0, 23]$$

$$\phi_g^{mid}(i, j, k) = \phi_g(i \times 2 + 18, j \times 2 + 18, k \times 2 + 18)/2 \\ i, j, k \in [0, 13]$$

$$\phi_g^{low}(i, j, k) = \phi_g(i \times 4 + 14, j \times 4 + 14, k \times 4 + 14)/4 \\ i, j, k \in [0, 8]$$

2. Gap filling: 为了提高预测精度, 在u型末端之间进一步插入额外的卷积层来连接低级特征和高级特征, 即图中绿色部分。

3. Multi-channel inputs: 差分图和梯度图也被用作网络的输入, 与原图像进行拼接。

◦ 从图像中抽出 $64 \times 64 \times 64$ 的patch作为输入, 输出 $24 \times 24 \times 24$  patch大小的形变场, 对应输入patch的中心区域。在对整个图像进行训练或应用网络时, 提取重叠的patch, 步长为24, 即输出patch大小。这样, 所有不重叠的输出patch就可以形成整个形变场。

### 3. 代码

无

### 4. 问题

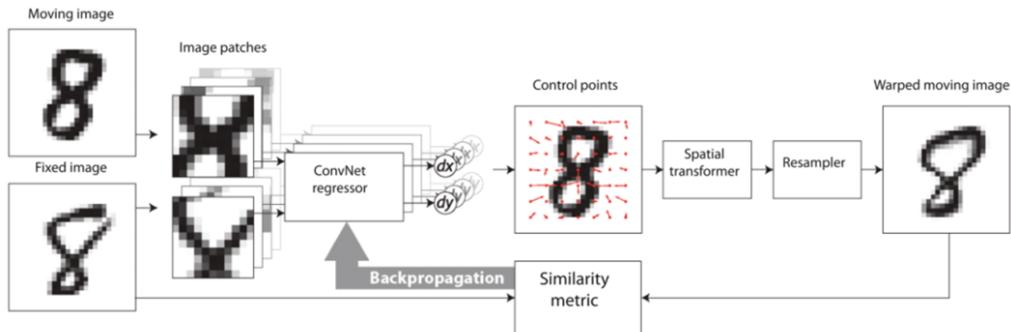
差分图和梯度图的理解和计算。

## 6. End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network (2017 DLMIA&ML-CDS&MICCAI DIRNet)

### 1. 动机

以前的方法中要么是传统的方法，要么是有监督的深度学习配准方法，作者提出了第一个无监督端到端的可形变的深度学习配准方法。

### 2. 方法



- ConvNet regressor是一个由Conv,Pool,BatchNorm组成的普通的神经网络，输出每个像素点在x和y方向的位移大小，然后经过一个STN网络得到配准图像。

### 3. 代码

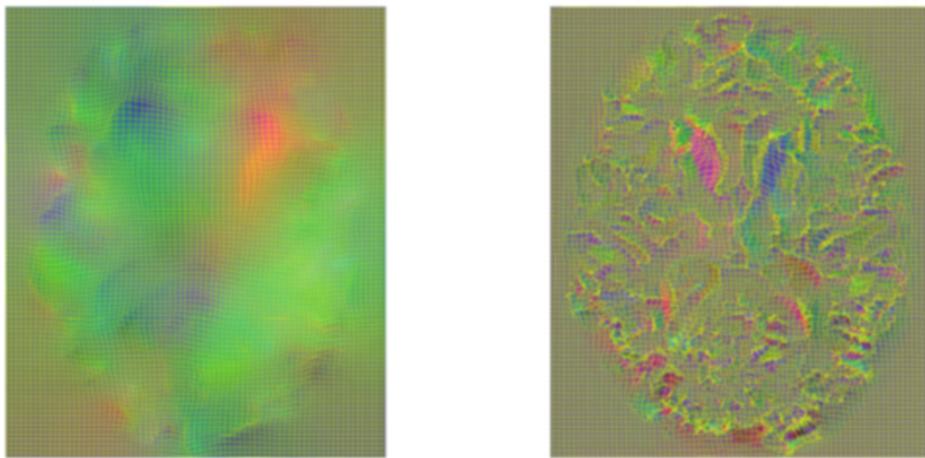
有，理解

### 4. 问题

无

## 7. Inverse-Consistent Deep Networks for Unsupervised Deformable Image Registration (2018 arxiv ICNet)

### 1. 动机



(a) Flow (not accurate), using a large weight for the smoothness constraint

(b) Flow (with folding), using a small weight for the smoothness constraint

Fig. 1. Illustration of two flows learned by a state-of-the-art deep-learning method [17] using (a) a large weight for the smoothness constraint and (b) a small weight for the smoothness constraint, respectively.

- 现有的大多数算法仅利用空间平滑惩罚来约束变换，这不能完全避免配准映射中的折叠（通常指示错误）。如果使用较大的权值作为平滑约束，过度鼓励待估计流动的局部平滑，如上图（a）所示，获得的配准结果会有全局错误。如果如果使用较小的权

值作为平滑约束，学习到的流中会出现大量的折叠，如图 (b) 所示，从而由于局部缺陷而产生错误配准。如何适当地调整平滑度约束的贡献，同时避免估计流量中的折叠，并保持较高的配准精度是很有挑战性。

- 以往的研究通常独立地估计从图像A到图像B或从图像B到图像A的变换，因此不能保证这些变换是彼此的逆映射，即忽略了一对图像之间转换的固有逆一致特性。

## 2. 方法

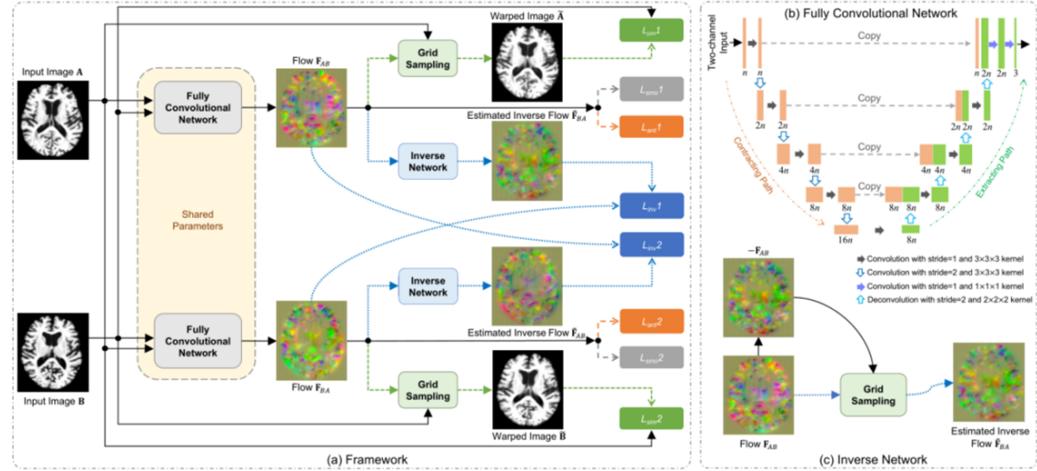


Fig. 2. Pipeline of the proposed Inverse-Consistent deep neural network (ICNet) for unsupervised deformable image registration, which takes a pair of images as input. (a) Framework of ICNet, (b) architecture of fully convolutional network (FCN), and (c) illustration of the inverse network. The term  $n$  in (b) denotes the number of starting convolutional channels in FCN.

- 为了解决这两个问题，本文提出了一种逆一致深度神经网络 (ICNet) 用于无监督变形图像配准。上图中两个Fully Convolutional Network实际上是同一个Network。
  - (a) 中绿色部分的是 $L_{sim}$ ，灰色部分是 $L_{smo}$ ，橙色部分是 $L_{ant}$ ，蓝色部分是 $L_{inv}$ 。 $F_{AB}$ 是将图像A配准到图像B的形变场。 $\tilde{F}_{BA}$ 是通过Inverse Network得到的形变场。在Inverse Network中，相当于把 $-F_{AB}$ 当做图像，用 $F_{AB}$ 对 $-F_{AB}$ 进行采样。
- 提出了一种反向一致约束 $L_{inv}$ ，以鼓励一对图像在多次传递中相互对称变形，直到双向变形的图像被匹配以实现正确的配准。

$$\mathcal{L}_{inv} = \| \mathbf{F}_{AB} - \tilde{\mathbf{F}}_{AB} \|_F^2 + \| \mathbf{F}_{BA} - \tilde{\mathbf{F}}_{BA} \|_F^2 \quad (2)$$

with

$$\begin{aligned} \tilde{\mathbf{F}}_{AB} &= \mathcal{G}(\mathbf{F}_{BA}, -\mathbf{F}_{BA}) \\ \tilde{\mathbf{F}}_{BA} &= \mathcal{G}(\mathbf{F}_{AB}, -\mathbf{F}_{AB}) \end{aligned} \quad (3)$$

where  $\mathcal{G}$  is the mapping generated by the grid sampling module (via a STN), and  $\| \cdot \|_F$  represents the Frobenius norm of a matrix. The two terms in Eq. 2 correspond to the notations  $L_{inv}1$  and  $L_{inv}2$  in Fig. 2 (a). By minimizing Eq. 2,

- 提出了一个反折叠约束 $L_{ant}$ ，以避免形变场发生折叠。

$$\begin{aligned} \mathcal{L}_{ant} = \sum_{p \in \Omega} \sum_{i \in \{x, y, z\}} \delta(\nabla \mathbf{F}_{AB}^i(p) + 1) &|\nabla \mathbf{F}_{AB}^i(p)|^2 \\ + \delta(\nabla \mathbf{F}_{BA}^i(p) + 1) &|\nabla \mathbf{F}_{BA}^i(p)|^2 \end{aligned} \quad (4)$$

where  $\nabla \mathbf{F}_{AB}^i(p)$  is the gradient of the flow  $\mathbf{F}_{AB}$  along the  $i$ -th ( $i \in \{x, y, z\}$ ) axis at the location of the voxel  $p$ . Besides, the term  $\delta(Q)$  is an index function used to penalize the gradient of the flow at the locations with foldings. That is, if  $Q \leq 0$ ,  $\delta(Q) = |Q|$ ; and  $\delta(Q) = 0$ , otherwise.

解释：以下图为例， $i$ 表示其中一个坐标轴方向（ $x$ ,  $y$ 或 $z$ ）， $m + 1$ 是 $m$ 在 $i$ 方向的相邻点， $F_{AB}^i(m)$ 是作用在点 $m$ 的 $i$ 方向上的位移， $m + F_{AB}^i(m)$ 表示点 $m$ 位移后的位置，其它符合类似。为了避免发生折叠，点 $m$ 和点 $m + 1$ 位移后形成的新的两个点应满足：

$$m + F_{AB}^i(m) < m + 1 + F_{AB}^i(m + 1) \Rightarrow F_{AB}^i(m + 1) - F_{AB}^i(m) + 1 > 0 \quad (1)$$

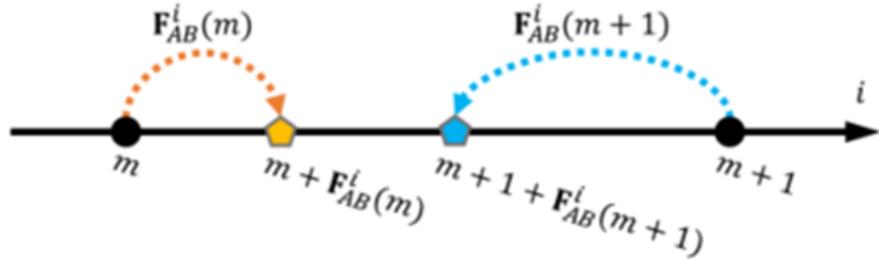
点 $m$ 在 $i$ 方向上的梯度定义为：

$$\begin{aligned} \nabla F_{AB}^i(m) &= \frac{F_{AB}^i(m + 1) - F_{AB}^i(m)}{(m + 1) - m} \\ &= F_{AB}^i(m + 1) - F_{AB}^i(m) \end{aligned} \quad (2)$$

结合公式 (1) 和 (2) 可以得到：

$$\nabla F_{AB}^i(m) + 1 > 0 \quad (3)$$

如果公式 (3) 则表示没有发生折叠，反之，在 $m$ 点发生折叠。



- 总的损失函数  $L = L_{smi} + \alpha L_{smo} + \beta L_{inv} + \gamma L_{ant}$ ，其中  
 $L_{smo} = \sum_{p \in \Omega} (\|\nabla F_{AB}(p)\|_2^2 + \|\nabla F_{BA}(p)\|_2^2)$ ,  
 $L_{smi} = (\|B - \tilde{A}\|_F^2 + \|A - \tilde{B}\|_F^2)$ ,  $\tilde{A}$ 和 $\tilde{B}$ 表示分别用用 $F_{AB}$ 和 $F_{BA}$ 配准后的图像。

3. 代码：

有。只看了损失函数部分，基本理解。

4. 问题

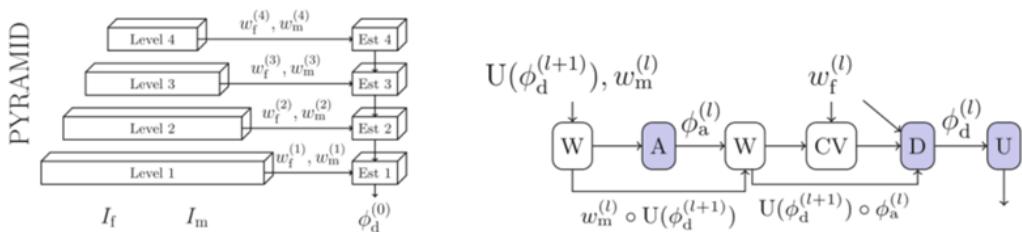
无

## 8. Learning a Deformable Registration Pyramid (2021 MICCAI)

1. 动机

提出了一种三维变形图像配准方法，灵感来自PWC-Net，一种流行于计算机视觉的二维光流估计的方法。

2. 方法



(a) Model architecture.

(b) Operations at each feature level.

- 图 (b) 中紫色部分表示有可训练参数，白色部分表示没有可训练参数。

- 四个层级输出四种不同大小的特征图，level 4的特征图最小。 $w_f^{(l)}$  是输入固定图像得到的特征图， $w_m^{(l)}$  是输入移动图像得到的特征图。从最上面的层开始每一层都要执行图 (b)。 $U(\phi^{(l+1)})$ 是上一层得到的形变场，初始值 $U(\phi^{(5)})$ 为全0的张量。**W**表示 warp操作，用得到的形变场 $U(\phi^{(l+1)})$  warp特征图 $w_m^{(l)}$ 。**A**表示放射变换网络，输入 $w_f^{(l)}$ 和**W**中得到的特征图，输出12个放射变换参数，上图可能少画了 $w_f^{(l)}$ 指向**W**的箭头。**CV**计算运动图像中的扭曲特征图与固定图像中的特征图之间的相关性，输出一个特征图。**D**是一个3D DenseNet，输入仿射变换后的特征图、**CV**得到的特征图和 $w_f^{(l)}$ ，输出形变场 $\phi^{(l)}$ 。重复这个过程。

- 损失函数：

$$\mathcal{L}_{\text{seg}}(S_f, S_m, \phi_d^{(0)}) = \lambda(1 - \text{DCS}(S_f, S_m \circ \phi_d^{(0)})), \quad (2a)$$

$$\mathcal{L}_{\text{sim}}^{(l)}(I_f^{(l)}, I_m^{(l)}, \phi_d^{(l)}) = -\gamma^{(l)} \text{NCC}(I_f^{(l)}, I_m^{(l)} \circ \phi_d^{(l)}), \quad (2b)$$

$$\mathcal{L}_{\text{smooth}}^{(l)}(\phi_a^{(l)}, \phi_d^{(l)}) = \alpha^{(l)} \|\phi_a^{(l)} - \phi_0^{(l)}\|_2^2 + \beta^{(l)} \|\nabla \phi_d^{(l)}\|_2^2, \quad (2c)$$

### 3. 代码：

有，浏览了一部分，不是很理解。

### 4. 问题：

**CV**模块的理解和计算没搞清楚。

## 9. Unsupervised 3D End-to-End Medical Image Registration with Volume Tweening Network (2019 JBHI VTN)

### 1. 动机

- 受FlowNet 2.0（一种用在光流估计中的网络）和STN的启发，作者提出了Volume Tweening Network (VTN)，它能够对端到端的CNN进行无监督训练，执行体素级3D医学图像配准。
- VTN包含了3个技术组件：（1）级联了注册子网络，这提高了注册大量移位图像的性能，并且没有太大的减速；（2）将仿射配准集成到我们的网络中，这被证明是有效的，比使用单独的工具更快；（3）在训练过程中加入了额外的可逆性损失，从而提高了配准性能。

### 2. 方法

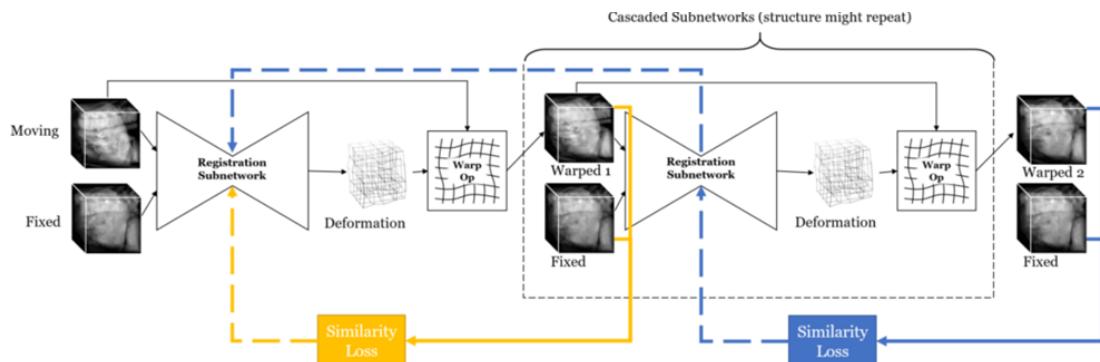


Figure 2: Illustration of the overall structure of Volume Tweening Network (VTN) and how gradients back-propagate. Every registration subnetwork is responsible for finding the deformation field between the fixed image and the current moving image. The moving image is repeatedly warped according to the deformation field and fed into the next level of cascaded subnetworks. The current moving images are compared against the fixed image for a similarity loss function to guide training. There is also regularization loss, but not drawn for the sake of cleanness. The number of cascaded subnetworks may vary; only two are illustrated here. In the figure, yellow (lighter) indicates the similarity loss between the first warped image and the fixed image, and blue (darker) indicates the similarity loss between the second warped image and the fixed image. Solid bold arrows indicate how the loss is computed, and dashed bold arrows indicate how gradients back-propagate. Note that the second loss will propagate gradients to the first subnetwork as a consequence of the first warped image being a differentiable function of the first subnetwork.

- 首先用一个FCN回归出12个仿射配准参数进行仿射变换，后面接n个级联的U-Net。黄色部分只向第一个子网络传递梯度，蓝色部分向前两个子网络传播梯度。除了级联多个网络，网络本身没有什么大的创新，提出了Invertibility Loss。

- Orthogonality Loss: 对于特定任务（医学图像配准），通常情况下，输入图像只需要小的缩放和旋转就可以仿射对齐。我们想对产生过度非刚性变换的网络进行惩罚。为此，我们引入了 $I + A$ 的非正交性的损失，其中 $I$ 表示单位矩阵， $A$ 表示仿射配准网络产生的变换矩阵（不包含平移项）。

$$L_{ortho} = -6 + \sum_{i=1}^3 (\lambda_i^2 + \lambda_i^{-2}) \quad (1)$$

其中， $\lambda_{1,2,3}$ 是 $I + A$ 的奇异值。如果 $A$ 具有很小的缩放和旋转，那么 $I + A$ 将接近与 $I$ ， $I$ 是正交的。当且仅当一个矩阵的所有奇异值都是1时，它是正交的。因此， $I + A$ 与正交矩阵的偏差越大(即奇异值偏离1越多)，其正交性损失越大。如果 $I + A$ 是正交的，其值将为0。

- Determinant Loss: 假设图像具有相同的手性，因此，不允许包含反射的仿射变换。这就要求 $\det(I + A) > 0$ 。结合正交性要求，设行列式损失为：

$$L_{det} = (-1 + \det(A + I))^2 \quad (2)$$

因为正交矩阵行列式为正负1，当行列式为-1时代表A存在反射变换？这时 $L_{det}$ 比较大。当 $I + A$ 不存在反射变换且满足正交性时 $L_{det}$ 接近0，否则会变大。

- Invertibility Loss: 不理解

straight arrow in Figure [5] Ideally, round-trip registration should satisfy the equations  $f_{12} * f_{21} = f_{21} * f_{12} = 0$ . We capture the round-tripness for a pair of images with the invertibility loss, namely

$$L_{inv} = \|f_{12} * f_{21}\|_2^2 + \|f_{21} * f_{12}\|_2^2. \quad (12)$$

The larger the invertibility loss, the less round-trip the registration. For perfectly round-trip registration, the invertibility loss is zero. We come up with, formulate, and implement the invertibility loss independently of [30]. We use L2 invertibility loss whereas [30] uses L1 left-right disparity consistency loss, which is just a matter of choice. We are the first to incorporate the invertibility loss into 3D images to boost performance on medical image tasks.

其中， $f_{12} * f_{21} = f_{21} + warp(f_{12}, f_{21})$ 。

### 3. 代码

无

### 4. 问题

Determinant Loss和Invertibility Loss不理解。没有明白网络的训练过程。

## 10. Recursive Cascaded Networks for Unsupervised Medical Image Registration (2019 ICCV)

### 1. 动机

- 一些研究也试图叠加多个网络。它们以非递归的方式为每个级联分配不同的任务和输入，并逐个训练它们，但它们的性能在只有少数（不超过3个）级联时接近极限。另一方面，级联在处理不连续和闭塞时可能没有多大帮助。因此，根据直觉，作者认为具有递归架构的级联网络适合可变形配准的设置。
- 然而，大多数提出的网络被强制进行简单的预测，这被证明是处理复杂变形时的负担，特别是大位移。DLIR和VTN也堆叠它们的网络，尽管它们都局限于少量的级联。DLIR一个接一个地训练每个级联，即在固定前级联的权重之后。VTN联合训练级联，而所有连续扭曲的图像都通过与固定图像的相似度来衡量。这两种训练方法都不允许中间级联逐步注册一对图像。这些非合作级联不考虑其他级联的存在而学习自己的目标，因此即使进行更多的级联，也很难实现进一步的改进。

- 级联方法已经涉及到计算机视觉的各个领域，例如级联分类器逐步改进了从监督训练数据中学习到的姿态估计，加快了目标检测的过程。
- 因此，作者提出递归级联体系结构，它鼓励对可以在现有基础网络上构建的无限数量的级联进行无监督训练，以提高技术水平。我们的体系结构与现有级联方法的不同之处在于，我们的每个级联通常将当前扭曲图像和固定图像作为输入，并且仅在最终扭曲图像上测量相似性（与DLIR, VTN相反），使所有级联能够协同学习渐进对齐。

## 2. 方法

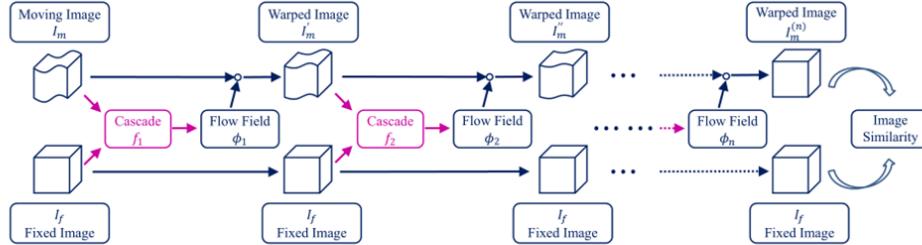


Figure 3. Illustration of our recursive cascade architecture. Circle denotes a composition, where the preceding warped image ( $I_m^{(k-1)}$ ) is reconstructed by the predicted flow field ( $\phi_k$ ), resulting in the successive warped image ( $I_m^{(k)}$ ). The unsupervised end-to-end learning is only guided by the image similarity between  $I_m^{(n)}$  and  $I_f$ , in contrast to previous works.

- 最终的预测可以被认为是由递归预测的流场的组合，而每个级联只需要学习一个简单的小位移对齐，可以通过更深的递归来细化，如下图所示。

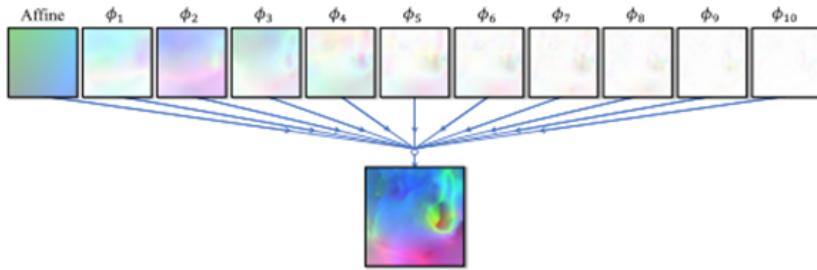


Figure 2. Composition of flow fields, corresponding to the example shown in Figure 1. The final flow prediction is composed of an initial affine transformation and  $\phi_1, \dots, \phi_n$ , each of which only performs a rather simple displacement. We can see that the top cascades mainly learn a global alignment, while the bottom cascades play a role of refinement. Flow fields are drawn by mapping the absolute value of the three components ( $x, y, z$ ) of flow displacements into color channels (R, G, B) respectively. White area indicates zero displacement.

- 每一个子网络都是类U-net网络。
- 在递归过程中可以重复应用一个级联，也就是说，多个级联可以使用相同的参数共享，这被称为共享权重级联。在每个级联之后立即插入一个或多个共享权重级联，即通过将每个  $f_k$  替换为  $n$  倍的  $f_k$  来构造总共  $r * n$  级联。这种方法在实验中被证明是有效的。当输出流场的质量可以通过进一步细化来提高时，测试过程中的共享权重级联是一种选择。然而，作者注意到这种技术并不总是得到积极的增益，并可能导致过度变形。递归级联只能确保扭曲的运动图像与固定图像之间的相似性不断增加，但如果图像过于完美匹配，聚合流场就会变得不那么自然。在训练中不使用共享权重级联的原因是，在使用的平台（Tensorflow）的梯度反向传播过程中，共享权重级联消耗的GPU内存与非共享权重级联一样大。要训练的级联的数量受到GPU内存的限制，但当数据集足够大以避免过拟合时，如果允许学习不同的参数，它们会表现得更好。
- 理论上，递归级联网络保持图像拓扑，只要每个子形变场都保持。然而，目前提出的方法中，折叠区域是常见的，并且在递归过程中可能会被放大，这给权值共享技术的使用带来了挑战。通过仔细研究正则化项，或设计一个保证可逆性的基本网络，可以减少这个问题。

### 3. 代码

有，没看

### 4. 问题

无

## 11. Affine Medical Image Registration with Coarse-to-Fine Vision Transformer (2022 CVPR C2FViT)

### 1. 动机

- 现有的基于CNN的仿射配准方法要么关注输入的局部错位，要么关注输入的全局方向和位置来预测仿射变换矩阵，对空间初始化敏感，脱离训练数据集的泛化能力有限。
- 在综合图像配准框架中，目标图像对通常在使用可变形(非刚性)配准之前基于刚性或仿射变换进行预对齐，消除了目标图像对之间可能的线性和大空间错位。
- 最近基于学习的变形图像配准方法的成功很大程度上是通过使用传统图像配准方法进行精确的仿射初始化，传统的配准方法具有较好的配准性能，但配准时间取决于输入图像之间的不对齐程度，对于高分辨率的3D图像体积，配准时间较长。
- 最近的一项研究[1]表明，**纯CNN编码器在一个看似微不足道的坐标变换问题中失败得很明显**，这意味着纯CNN编码器可能不是一个理想的架构，用于编码笛卡尔空间中图像扫描的方向和绝对位置或仿射参数。[2]还报告了基于CNN的仿射配准方法在实践中表现不佳，即使是对于具有较大接受野的深度CNN。
- 受到最近Vision Transformer模型成功的激励，作者脱离了现有的基于cnn的方法，提出了一种用于3D医学仿射注册的粗到细视觉转换器(C2FViT)。这是第一个基于学习的仿射配准方法，在学习三维医学图像配准的全局仿射配准时考虑输入图像之间的非局部依赖性。
- 主要贡献： (1) 定量研究和分析了现有基于学习的仿射配准方法和传统仿射配准方法在三维大脑配准中的配准性能、鲁棒性和泛化性。 (2) 提出了一种新的基于学习的仿射配准算法，即C2FViT，该算法利用卷积视觉转换器和多分辨率策略。C2FViT优于最近的基于CNN的仿射配准方法，同时在数据集上表现出优越的鲁棒性和泛化性； (3) 所提出的学习范式和目标函数可以适应各种参数配准方法。

### 2. 方法

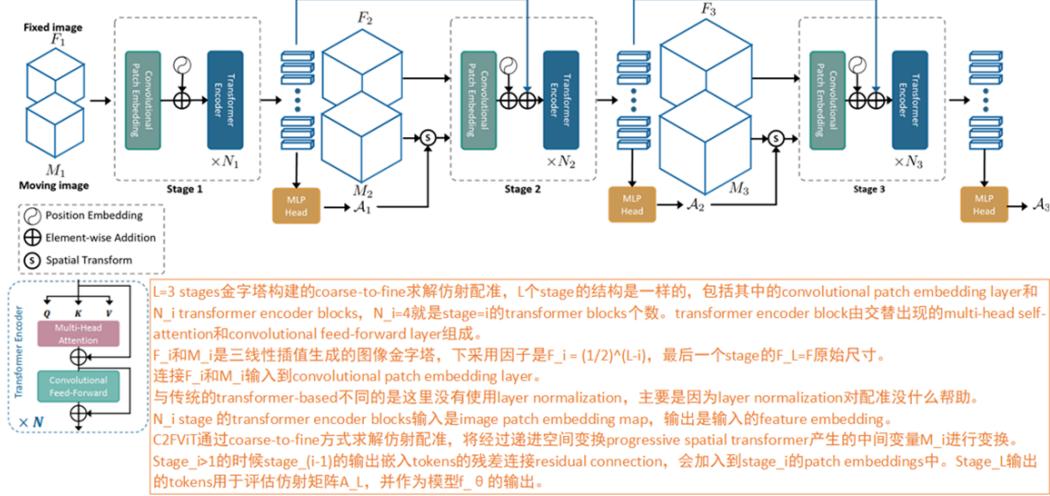
卷积方法求解仿射配准迭代优化一般都是迭代优化问题，优化方法一般使用adaptive gradient descent或convex optimization。但是配准时间与输入图像对的复杂程度和分辨率相关。CNN的做仿射配准主要有两类：concatenation-based和Siamese network。

concatenation-based仿射子网络将连接的fixed和moving作为输入，并使用single-stream CNNs提取数据的局部错位local misalignment。仿射配准是全局的，但是concatenation-based配准网络缺乏全局联通，只关注两个图像空间之间重叠区域，无法处理初始偏差较大的情况。

Siamese Network对fixed和moving分别提取特征，并在管道末端使用全局平均池化global average pooling，增强网络的全局方向和仿射变换进行编码。虽然能够捕获全局的高级几何全局high-level geometrical，忽略了输入图像之间初始不对齐的局部特征。

有研究表明，纯CNN编码器在坐标变化问题上表现不如，纯CNN编码器不适合编码Cartesian空间或仿射参数，即便是再大的感知野也不行。预配准如果配的不准确的话会影响配准精度或者形变配准算法的收敛。

CNNs由于固有归纳偏置inductive biases的限制（权重共享和局部性），在远程依赖性long-range方面有局限性。ViT提供较少的图像特异性归纳偏置inductive biases，在大规模数据集训练时有巨大的潜力。ViT是一个纯transformer构建的金字塔结构，用以模仿CNN中多尺度策略。ViT中引入适度的convolutional inductive bias可以提高整体性能，特别是对于小数据集。虽然CNN在形变配准中取得较大的成功，但作者依旧认为cnn并不适合仿射配准。因为仿射配准通常是为了检出较大的线性偏移，这是一种全局的操作，这就与CNN的inductive bias矛盾。



由于self-attention机制使得ViT对sequence of non-overlapping image patches的long-range dependencies建模能力突出，但是vision transformer缺乏locality机制，也就是很难对相邻patch进行建模。

C2FViT中给transformers增加了locality，主要是通过patch embedding和feed-forward layer。

相比传统的transformer，将linear patch embedding替换成convolutional patch embedding。Convolutional patch embedding layer的目的是将输入图像转到sequence of overlapping patch embeddings。

输入 $I \in R^{(H \times W \times D \times C)}$ ，其中I的空间维度是H、W和D，而C是channels。Convolutional patch embedding layer利用3D convolutional layer计算输入I的patch embedding map  $Z \in R(H_i \times W_i \times D_i \times d)$ 。这个3D convolutional layer的kernel size = k^3、stride = s、padding = p、channel=d。

然后将patch embedding map Z平展flatten成一个sequence of patch embeddings(tokens)  $[Z^i]_{i \in R_d}$   $i=1, \dots, N$ ，其中  $N = H_i \times W_i \times D_i \times d$  是embedding dimension。然后将这些patch embedding  $Z^i$ 聚合成matrix  $Z^i \in R^{(N \times d)}$ 。其中  $N=4096$ ,  $d=256$ ,  $S=(H/16, W/16, D/16)$ 。为了强制window overlapping将卷积操作的sliding window大小设置为  $s = 2s - 1$ ，然后进行0填充  $p=[k/2]$ 。与ViT中的linear patch embedding相比C2FViT中的convolutional patch embedding有助于建模fixed和moving的local spatial context 和 feature。同时可以调整patch embedding的number和feature dimension。

Transformer encoder中feed-forward layer是唯一的local和translation equivariance，feed-forward layer包含MLP和两个隐藏层。由于ViT的feed-forward layer是以patch-wise方式应用到patch embedding map，它缺乏一个local机制对相邻patch embedding之间的关系进行建模。

C2FViT的feed-forward中的MLP block包含你两个hidden layers，在两个hidden layers之间包含一个 $3 \times 3 \times 3$ 的depth-wise convolution layer。Depth-wise convolution进一步将locality加入到C2FViT的transformer中。

transformers由于self-attention是的他们在sequence of embedding的long-range dependencies建模方面非常擅长。与基于CNN的仿射配准方法相比，通过C2FViT的transformer encoder投影的query-key pairs间的亲相似性。

$$Q = \hat{Z}W^Q, K = \hat{Z}W^K \text{ and } V = \hat{Z}W^V$$

Query Q、Key K和Value V都是patch embeddings (tokens) 线性投影得到的。进一步将self-attention转换成multi-head self-attention(MHA)，attention heads个数为h=2，attention head j的线性投影矩阵 $W^Q_j$ 、 $W^K_j$ 、 $W^V_j \in R^{(d \times d_h)}$ ， $d_h=d/h$ 大小是一样的， $d_h$ 是attention head 的嵌入维度，attention head j的attention操作是：

$$\text{Attention}(Q_j, K_j, V_j) = \text{Softmax}\left(\frac{Q_j K_j^T}{\sqrt{d_h}}\right) V_j \quad (1)$$

最后将所有的attention heads通过矩阵 $W^O \in R^{d \times d}$ 进行连接和线性投影。

C2FViT采取多分辨率策略。C2FViT的每个stage的末尾都有一个classification head，该classification head由两个连续的多层感知机two successive multilayer perceptrons(MLP) layer组成，这两层MLP使用的激活函数是双曲正切hyperbolic tangent(Tanh)。Classification head输入为averaged patch-wise patch embedding，输出为仿射变换参数。中间阶段stage  $i$ 输出的仿射矩阵作用在moving图像 $M_{-i+1}$ ，得到变换后的图像 $M_{-i+1}$ 和 $F_{-i+1}$ 拼接，输入到stage  $i+1$ 。提出来的递减空间变换、线性错位linear misalignment很容易在低分辨率输入下被消除，Higher leave的transformer用来消除复杂的misalignment。在higher stages降低问题的复杂度。

以往的仿射配准算法输出的仿射矩阵不能分解为一组线性的几何变换：translation、rotation、scaling和shearing。C2FViT预测第 $i$ 几何变换参数，并不是直接预测仿射矩阵。仿射配准问题就衰减为 $\theta(F, M) = [t, r, s, h]$ ，这里的 $t, r, s, h \in \mathbb{R}^4$ 分别表示translation、rotation、scaling和shearing参数。通过矩阵乘可以得到仿射矩阵 $A = T \times R \times S \times H$ 。通过修改或裁剪不需要的几何变换，C2FViT可以很容易的其他的变换，比如删除scaling和shearing矩阵，我们就得到了刚体变换。另外变换模型有几何约束，这样就减少了优化过程中的搜索空间。Rotation和shearing  $\in [-\pi, \pi]$ ；translation  $\in [-50%, 50%]$  × 最大的分辨率，scaling  $\in [0.5, 1.5]$ ，C2FViT的rotation和shearing不是几何中心而是质心c\_l。如果说图像背景强度不是0的话，就应该设置成几何中心。

$$c_I = \frac{\sum_{p \in \Omega} p I(p)}{\sum_{p \in \Omega} I(p)}$$

$$I_\theta(F, M) = A_f$$

$I_\theta$ 是C2FViT模型， $A_f$ 是放射变换矩阵。模型的目标就是最小化方程，其中 $\theta$ 就是C2FViT的学习参数，fixed和moving从数据集中随机采样，L评估图像的便宜程度

$$\theta^* = \arg \min_{\theta} \left[ \mathbb{E}_{(F, M) \in D} \mathcal{L}(F, M(\phi(A_f))) \right], \quad (2)$$

无监督学习中损失函数用的是negative NCC similarity

$$\mathcal{L}_{sim}(F, M(\phi)) = \sum_{i \in [1..L]} -\frac{1}{2^{(L-i)}} \text{NCC}_w(F_i, M_i(\phi)), \quad (3)$$

这里的L是金字塔的等级，NCC\_w表示windows size  $w^3$ 的local normalized cross-correlation

半监督学习就是在里面加入了segmentation

$$\mathcal{L}_{seg}(S_F, S_M(\phi)) = \frac{1}{K} \sum_{i \in [1..K]} \left( 1 - \frac{2(S_F^i \cap S_M^i(\phi))}{|S_F^i| + |S_M^i(\phi)|} \right) \quad (4)$$

K表示label的个数，文章中L=3,  $\lambda = 0.5$

一般的仿射配准方法都是通过质心对齐center of mass(CoM)得到的平移参数进行初始配准的。作者也对是否使用CoM配准进行评估

数据处理的时候重采样使用的是trilinear interpolation。评估的时候用的是完整的分辨率，Adam优化器使用固定的learning rate=1e^-4，batch size =1

from <https://blog.csdn.net/fanre/article/details/124971244>

- 图中， $F_1, F_2, F_3$ 分别表示不同分辨率大小的固定图像。

### 3. 代码

有，没看

### 4. 问题

CoM的理解

- [1] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. arXiv preprint arXiv:1807.03247, 2018.
- [2] Zhengyang Shen, Xu Han, Zhenlin Xu, and Marc Niethammer. Networks for joint affine and non-parametric image registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4224–4233, 2019. 1, 2.

## 12. Pyramid Vision Transformer A Versatile Backbone for Dense Prediction without Convolutions (2021 ICCV PVT)

### 1. 动机

- 旨在探索一种超越CNN的替代骨干网，可用于密集预测任务，如目标检测和分割等。
- 虽然ViT适用于图像分类，但直接将其应用于像素级的密集预测，如物体检测和分割，具有挑战性，因为（1）它的输出特征图是single-scale和低分辨率的，（2）它的计算和内存成本相对较高。
- 为了解决上述限制，这项工作提出了一个纯Transformer骨干网，称为 Pyramid Vision Transformer (PVT)，它可以在许多下游任务中作为CNN骨干网的替代方案，包括图像级预测和像素级密集预测。具体而言，PVT克服了传统Transformer的困难：（1）将细粒度图像补丁（即每个补丁 $4 \times 4$ 像素）作为输入来学习高分辨率表示，这对于密集预测任务是必不可少的；（2）引入渐进式收缩金字塔，随着网络的深入减少Transformer的序列长度，显著降低计算成本；（3）采用空间约简注意（space-reduction attention, SRA）层，进一步减少学习高分辨率特征时的资源消耗。

## 2. 方法

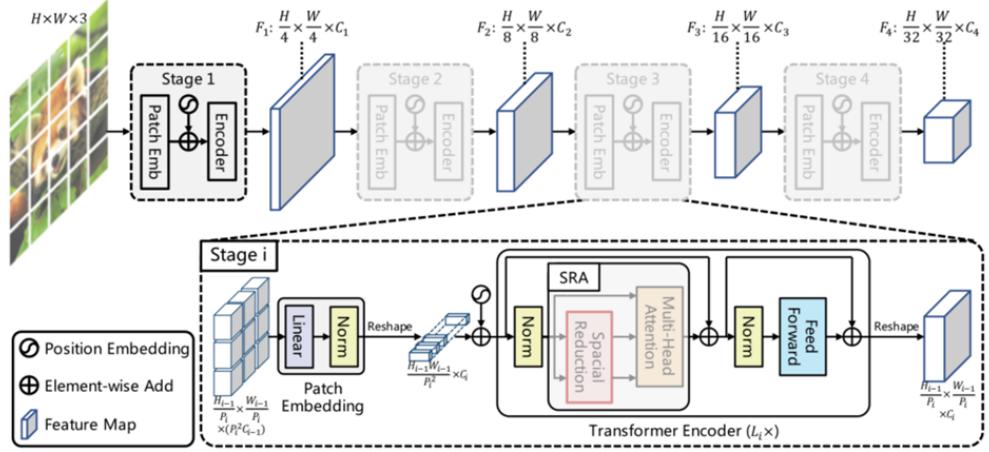
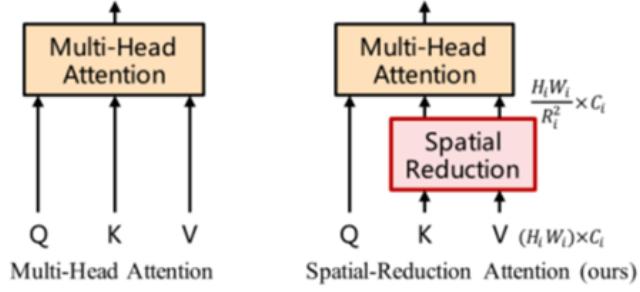


Figure 3: **Overall architecture of Pyramid Vision Transformer (PVT)**. The entire model is divided into four stages, each of which is comprised of a patch embedding layer and a  $L_i$ -layer Transformer encoder. Following a pyramid structure, the output resolution of the four stages progressively shrinks from high (4-stride) to low (32-stride).

- PVT采用逐级收缩策略通过patch embedding控制特征图的尺度。
- PVT为了进一步减少计算量，将常规的multi-head attention (MHA) 用spatial-reduction attention (SRA)来替换。SRA的核心是减少attention层的key和value对的数量，常规的MHA在attention层计算时key和value对的数量为sequence的长度，但是SRA将其降低为原来的 $1/2$ 。SRA的具体结构如下所示：



输入  $x \in \mathbb{R}^{(H_i W_i) * C_i}$ ，然后将它reshape到  $\frac{H_i W_i}{R_i^2} * R_i^2 C_i$ ，然后再经过一个线性投影  $W_S \in \mathbb{R}^{(R_i^2 C_i) * C_i}$ ，尺寸就变成了  $\frac{(H_i W_i)}{R_i^2} * C_i$ ，这样  $K$  和  $V$  的长度变小了  $\frac{1}{R_i^2}$ ，计算量也就变为了原来的  $\frac{1}{R_i^2}$  倍。

- 其他的基本和ViT基本一样。
- 总体来说，PVT的优点有：
  - (1) 传统CNN backbone的感受野随着深度增加而逐渐增大，而PVT始终保持全局感受野（受益于Transformer中的自注意力机制，在所有patchs中执行注意力），这对检测、分割任务更为合适；
  - (2) 相比ViT，引入了金字塔结构的PVT可以嵌入到很多经典的pipelines中，如 RetinaNet、Mask-RCNN等；
  - (3) 可以和其他Transformer Decoder结合，组成无卷积的框架，如PVT+DETR进行端到端目标检测。

## 3. 代码

有，没看

## 4. 问题

无

## 13. TransMorph: Transformer for unsupervised medical image registration (2021 Media TransMorph)

### 1. 动机

- 由于卷积运算的固有局域性（即有限的接受域），卷积网络架构通常在建模图像中显示的远程空间关系（即彼此相距较远的两个体素之间的关系）时存在局限性。U-Net通过在ConvNet中引入向下和向上采样操作来克服这一限制，理论上扩大了ConvNet的接受域，从而鼓励网络考虑图像中点之间的远程关系。然而，仍然存在几个问题：第一，前几层的接受域仍然受到卷积核大小的限制，图像的全局信息只能在网络的较深层查看；其次，研究表明，随着卷积层的加深，来自遥远体素的影响会迅速衰减。因此，在实践中，U-Net的有效感受野比它的理论感受野要小得多。这限制了U-Net感知语义信息和建模点之间长期关系的能力。然而，人们认为理解语义场景信息的能力在应对大变形方面非常重要。
- Transformer可以是图像配准的一个强有力候选者，因为它可以更好地理解移动和固定图像之间的空间对应关系。配准是建立这种对应关系的过程，直观地，通过比较运动和固定图像的不同部分。ConvNet的感受野较窄：它在局部进行卷积，其感受野随ConvNet的深度成比例增长；因此，浅层有一个相对较小的感受野，限制了ConvNet将两幅图像之间的远处部分联系起来的能力。例如，如果运动图像的左边部分与固定图像的右边部分相匹配，如果ConvNet不能同时看到这两个部分（即当其中一个部分落在ConvNet的感受野之外），它将无法在这两个部分之间建立适当的空间对应关系。而Transformer由于其较大的感受野和自注意机制，能够处理这种情况，并迅速聚焦到需要变形的部分。
- 这篇文章的主要贡献如下：
  - (1) Transformer-based model：介绍了利用Transformer进行图像配准的开创性工作。提出了一种新的基于transformer的神经网络TransMorph，用于仿射和变形图像配准。
  - (2) Architecture analysis：实验证明，位置嵌入是Transformer中常用的元素，对于所提出的Transformer-convnet混合模型来说，位置嵌入是不需要的。其次，实验证明基于transformer的模型具有比ConvNets更大的有效接受场。此外，证明了TransMorph促进一个平坦的配准损失landscape。
  - (3) Diffeomorphic registration：证明了TransMorph可以很容易地集成到两个现有的框架中，作为一个注册骨干来提供同胚配准。
  - (4) Uncertainty quantification：提供了TransMorph的贝叶斯不确定性变体，产生transformer不确定性和完美校准的appearance不确定性估计。

### 2. 方法

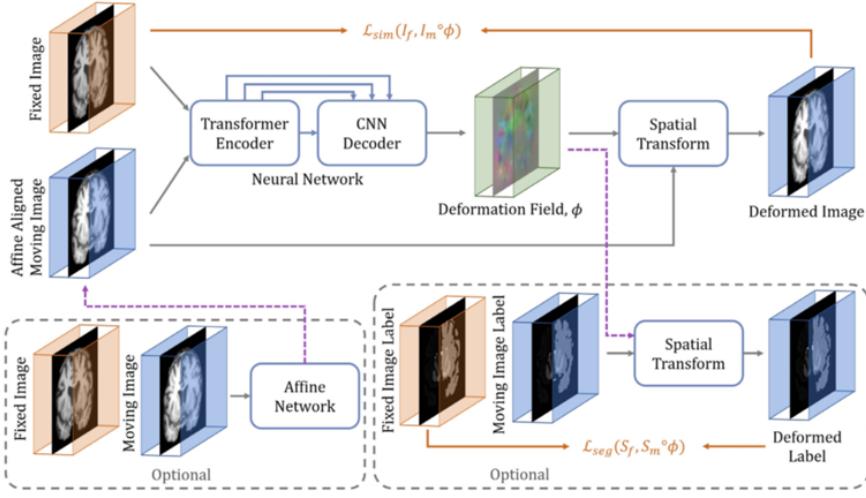


Fig. 3. The overall framework of the proposed Transformer-based image registration model, TransMorph. The proposed hybrid Transformer-ConvNet network takes two inputs: a fixed image and a moving image that is affinely aligned with the fixed image. The network generates a nonlinear warping function, which is then applied to the moving image through a spatial transformation function. If an image pair has not been affinely aligned, an affine Transformer may be used prior to the deformable registration (left dashed box). Additionally, auxiliary anatomical segmentations may be leveraged during training the proposed network (right dashed box).

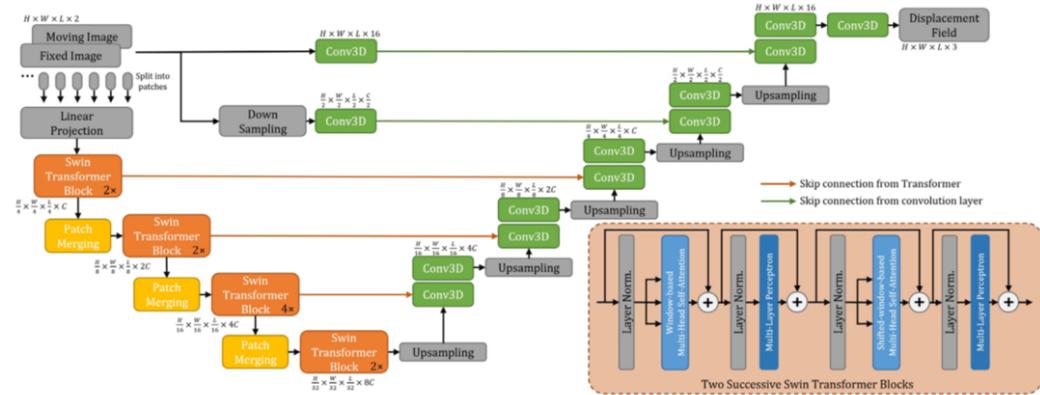


Fig. 1. The architecture of the proposed TransMorph registration network.

- 仿射配准网络使用一个Tranformer，回归出旋转、平移、缩放和剪切参数，是一个单独拿出来训练的网络。
- 下面那个图是Encoder和Decoder详细的结构。

### 3. 代码

有，没看

### 4. 问题

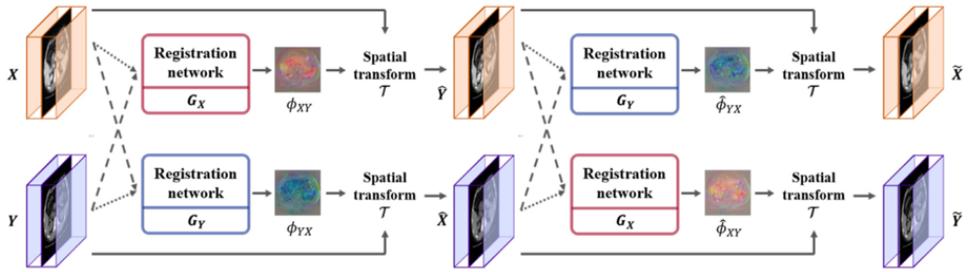
无

## 14. CycleMorph: Cycle consistent unsupervised deformable image registration (2020 Media CycleMorph)

### 1. 动机

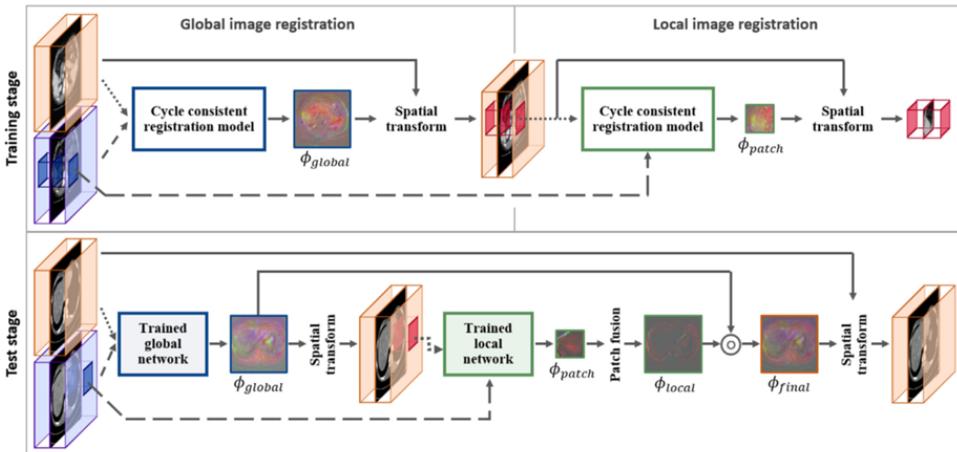
- 现有的深度学习方法在具有配准向量场的变形过程中对原始拓扑结构的保留方面存在一定的局限性。为了解决这个问题，作者提出了一种循环一致的可变形图像配准方法。
- 另一个重要创新是扩展到多尺度实现，以处理大容量图像配准。

### 2. 方法



**Fig. 2.** The overall framework of the proposed cycle consistent deep learning model, CycleMorph, for deformable image registration. Two registration networks ( $G_X, G_Y$ ) are used to take inputs by switching their orders. Each network takes two volumes ( $X, Y$ ) and computes displacement vector fields. Short and long dashed lines denote the moving images and fixed images, respectively. The spatial transform function deforms the moving image according to the vector fields to match a shape of the fixed image. These transformed images ( $\tilde{X}, \tilde{Y}$ ) are taken to the networks followed by transform function to ensure that the deformed images can be returned to original state.

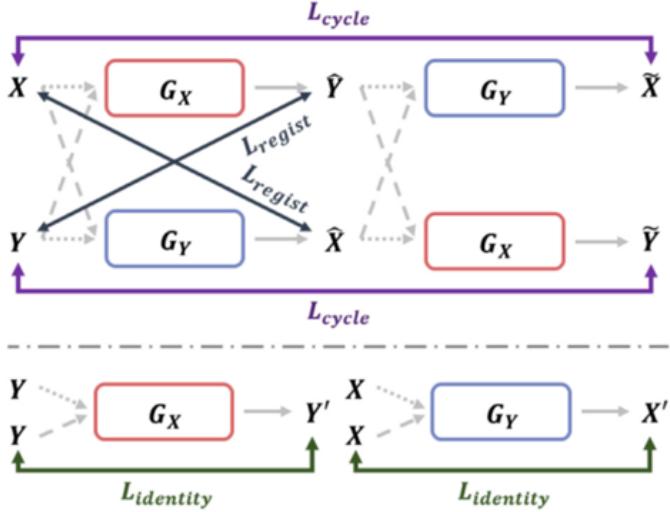
- 配准网络  $G_X : (X, Y) \rightarrow \phi_{XY}$ , 其中  $\phi_{XY}$  表示从  $X$  到  $Y$  的形变场;  $G_Y$  同理。
- 图中短虚线和长虚线分别表示移动图像和固定图像。
- 为了保证变形图像与固定图像之间的拓扑保持, CycleMorph 采用了原始运动图像与再变形图像之间的循环一致性约束。也就是说, 这两个变形的图像被再次作为网络的输入, 通过切换它们的顺序来对图像的像素级施加循环一致性。这种约束源于一个数学观察, 即两个拓扑空间之间的同胚映射保留了所有的拓扑性质, 但与需要在网络实现方面可微映射的附加条件的微分同胚映射相比, 它需要更宽松的约束。由于微分同胚是同胚变形的一个子集, 通过保证变形图像的形状连续恢复到原来的形状, 使网络具有循环一致性, 使网络能够提供能够保持拓扑的同胚映射。



**Fig. 4.** Flow diagram of the multiscale CycleMorph registration method for large volume images. The upper part illustrates the flow of training stage for global and local image registration. The lower part shows the flow of test stage using the trained global and local registration networks for a given moving and fixed images. The short- and long-dashed lines indicate moving image and fixed image, respectively.

- 在测试阶段, 运动图像的连续变形是由局部配准网络控制的, 两种配准网络由于每个阶段插补误差的累积而潜在地降低了配准精度。因此, 不是对运动图像进行两次变形, 而是将训练好的全局和局部网络依次应用于每个尺度上的变形场估计, 并且使用细化的变形场仅对运动图像进行一次最终变形。具体来说, 给定一对由运动图像和固定图像组成的新输入, 经过训练的全局配准网络生成中间变形图像和相应的变形向量场  $\phi_{global}$ 。然后, 局部配准网络以变形图像和固定图像中提取的patch为输入, 为每个patch生成变形场  $\phi_{patch}$ 。然后, 对局部网络生成的变形场进行拼接, 并按一定的间隔进行重叠, 得到精细尺度上  $\phi_{local}$  的局部变形场。在这里, 为了得到在每个patch的边界处形成合理变形的位移, 重叠大小设置为较大。然后, 由全局变形域和局部变形域的组合  $\phi_{global} \circ \phi_{local}$ , 得到最终的变形向量域  $\phi_{final}$ 。这种合成方法可以用  $\phi_{local}$  扭曲  $\phi_{global}$ , 并将结果与  $\phi_{local}$  相加来实现 [1]。最后, 利用  $\phi_{final}$  和空间变换器对运动图像进行一次变形, 使其与固定目标图像对齐。

[1] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, Nicholas Ayache.  
Diffeomorphic demons: Efficient non-parametric image registration,  
NeuroImage, 45 (1), S61-S72 2009.



**Fig. 3.** The diagram of loss function structure in our proposed method. The registration loss function,  $\mathcal{L}_{register}$ , computes dissimilarity in shape of the deformed and fixed images. The cycle loss function,  $\mathcal{L}_{cycle}$ , allows the displacement fields to preserve topology between the moving and deformed images. The identity loss function,  $\mathcal{L}_{identity}$ , prevents the network from generating deformation fields that deform stationary regions.

- 损失函数:

$$L = \mathcal{L}_{register}(X, Y, G_X) + \mathcal{L}_{register}(Y, X, G_Y) + \alpha \mathcal{L}_{cycle}(X, Y, G_X, G_Y) + \beta \mathcal{L}_{identity}(X, Y, G_X,$$

- $\mathcal{L}_{register} = -(T(X, \phi_{XY}) \otimes Y) + \lambda \sum \| \nabla \phi_{XY} \|^2$  表示局部归一化互相关操作。

- 为了保留移动图像和变形图像之间的拓扑结构，在图像的像素级上设计了循环一致性，如图3所示。具体来说，图像  $X$  首先变形为图像  $Y$ ，之后，变形的图像再次被另一个网络配准，生成图像  $\bar{X}$ 。然后，循环一致性应用于重新变形的图像  $\bar{X}$  与其原始图像  $X$  之间，使得  $\bar{X} \simeq X$ 。

$$\mathcal{L}_{cycle}(X, Y, G_X, G_Y) = \| T(\hat{Y}, \hat{\phi}_{XY}) - X \|_1 + \| T(\hat{X}, \hat{\phi}_{YX}) - Y \|_1.$$

- 当通过位移矢量场对图像进行变形时，图像的静止区域不应更改为固定点。考虑到这一点并提高配准精度，如图 3 所示，通过强加输入图像在相同图像用作移动图像和固定图像时不应变形来设计恒等约束。

$$\mathcal{L}_{identity}(X, Y, G_X, G_Y) = -(T(Y, G_X(Y, Y)) \otimes Y) + -(T(X, G_Y(X, X)) \otimes X)$$

### 3. 代码

有，没看

### 4. 问题

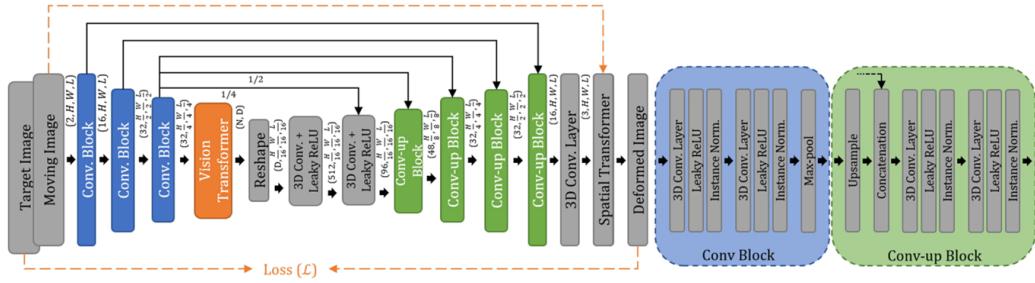
无

## 15. ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration (2021 arxiv ViT-V-Net)

### 1. 动机

- 尽管卷积网络的性能很有希望，但由于卷积运算的固有局域性，卷积网络架构通常在建模图像中显式的远程空间关系（即彼此相距很远的两个体素之间的关系）时存在局限性。近年来，由于基于自注意力的体系结构在自然语言处理方面的巨大成功。
- 在这项工作中，作者提出了第一个研究，以调查使用ViT的三维医学图像配准。我们提出了采用混合ConvNet-Transformer架构进行自监督体积图像配准的ViT-V-Net。在该方法中，ViT被应用于移动和固定图像的高级特征，这需要网络学习图像中点之间的远距离关系。编码器和解码器级之间使用长跳接来保持定位信息的流动。

### 2. 方法



**Figure 1:** Method overview and network architecture of ViT-V-Net.

- 没什么特点

### 3. 代码

有，没看

### 4. 问题

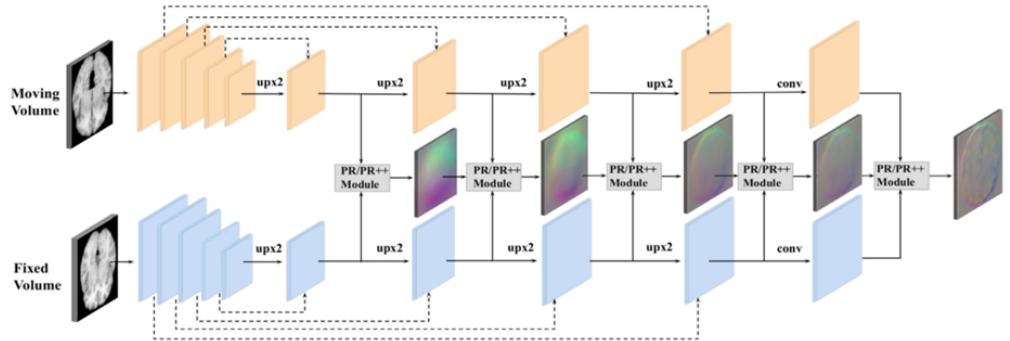
无

## 16. Dual-stream pyramid registration network (2020 MICCAI/Media Dual-PRNet)

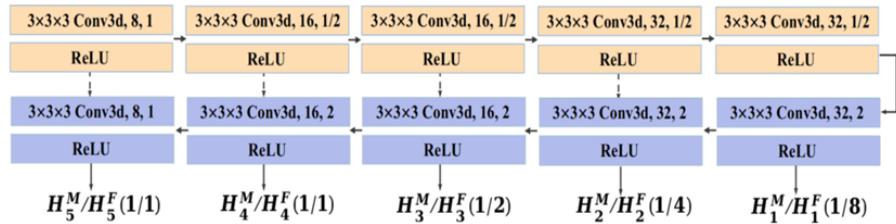
### 1. 动机

- Lewis等人证明[1]，现有基于CNN的方法的性能在现实世界的临床应用中可能受到限制，其中两个医学图像或体积可能具有显著的空间位移或大的切片空间。
- 最近的光流估计方法试图通过逐渐细化估计流来处理大位移。这启发作者设计了一种连续warp机制，能够以从粗到细的方式逐步warp两个卷。
- 除了学习有意义的特征表示外，医学图像配准还需要在移动和固定体积之间具有很强的像素级对应关系，这自然涉及到学习移动和固定体积中间特征之间的局部相关性。目前的光流估计方法利用关联层使网络能够从卷积特征中识别实际对应关系。这也促使作者开发了一种新的3D相关层，能够学习这种相关性，以进一步增强特征表示。
- 基于CNN的方法可能无法估计复杂变形场中的大位移，最近的工作已致力于通过开发堆叠多网络来解决这一问题，DLIR, CycleMorph, Cursive Cascaded等。然而，多个网络的顺序组合将导致插值伪影（artifacts）的累积，这可能会影响估计变形场的质量。因此，最近的方法试图在多个分辨率下估计变形场。然后介绍了双流编码器和金字塔框架。

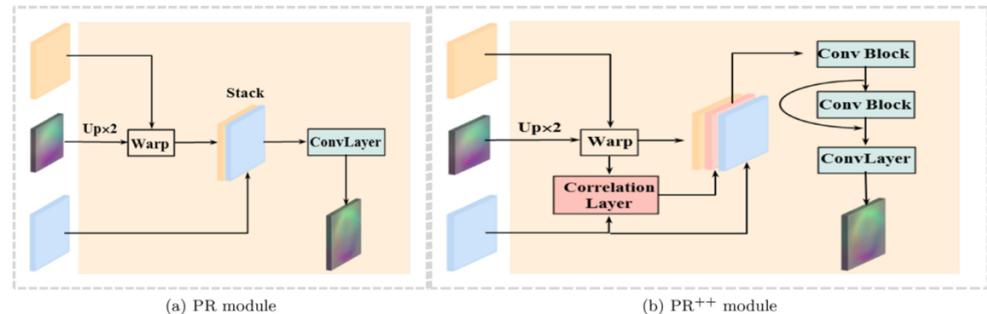
### 2. 方法



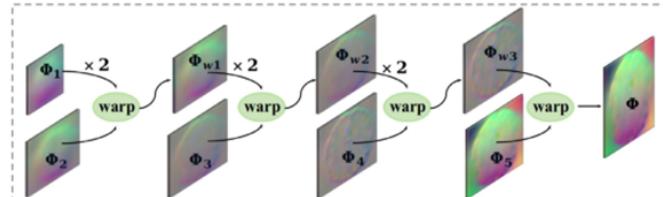
**Fig. 1.** Framework of the proposed Dual-PRNet<sup>++</sup>, which is a dual-stream encoder-decoder network, integrated with new sequential pyramid registration including a sequence of pyramid registration (PR) modules or PR<sup>++</sup> modules. The dual-stream model computes two convolutional feature pyramids separately from two input volumes, while the PR / PR<sup>++</sup> modules estimate a sequence of deformation fields which can warp the pyramid features gradually in a coarse-to-fine manner. Finally, the final deformation field is generated by sequentially warping the estimated fields, as shown in Fig. 4.



**Fig. 2.** Backbone of the proposed Dual-PRNet<sup>++</sup>. It consists of an encoder (yellow) and a decoder (blue), each of which has five convolutional blocks. The convolutional layers are indicated by the filter size, the number of output channels, and spatial resolution (w.r.t. the feature maps of previous layer).  $H_i^M$  and  $H_i^F$  denote the feature maps computed from the moving volume and the fixed volume separately, with various spatial resolutions (w.r.t. the input volumes) at different convolutional blocks.



**Fig. 3.** The proposed (a) Pyramid Registration (PR) module, and (b) its extension: PR<sup>++</sup> module, which improves the PR module by computing correlation features with further enhancement by residual convolutions.



**Fig. 4.** The final deformation field is computed by sequentially warping the current field with the previous one ( $\times 2$  up-sampling).

- 两个输入图像分别计算两个卷积特征金字塔，生成更强的变形估计深度特征 (Figure 2)。
- 提出顺序金字塔配准，其中配准场序列由一组设计的金字塔配准(PR)模块估计。估计的配准字段在解码层上执行顺序的扭曲，以从粗到细的方式逐渐细化特征金字塔。这使模型具有较强的处理大变形的能力 (Figure 1, Figure 3)。
- PR模块可以通过计算两个特征金字塔之间的局部3D相关性，然后进行多个残差卷积来进一步增强，这些残差卷积聚集了更丰富的解剖结构局部细节，从而更好地估计变形场 (Figure 3)。
- 设计了三维相关层来计算卷积特征空间中两个输入体之间的局部相关性。这使我们能够聚合相关的特征，这些特征在原始PR模块中没有直接探索，但可以在深度表示中强调局部细节。

Specifically, let  $P_i^W$  and  $P_j^F$  denote the central voxel of the 3D blocks (with size of  $(2k+1)^3$ ) sampled from the feature maps of the warped moving volume and the fixed volume. The correlation relationship between the two sampled 3D blocks can be computed as:

$$C(w_i, f_j) = \frac{1}{(2k+1)^3} \sum_{n_w, n_f \in [-k, k]^3} p_{i+n_w}^W \times p_{j+n_f}^F \quad (4)$$

where  $n \in [-k, k]^3$  means  $n$  iterates over a 3D neighborhood  $[-k, k] \times [-k, k] \times [-k, k]$  of  $P_i^W$  or  $P_j^F$ . In our experiments,  $k$  was empirically set to 1. Given a local 3D block on the feature maps of

- 模型生成5个分辨率递增的连续变形字段:  $[\Phi_1, \Phi_2, \Phi_3, \Phi_4, \Phi_5]$ 。为了计算最终的变形场, 将估计的变形场上采样2倍, 然后用接下来估计的变形场进行扭曲。这样的上采样和扭曲操作是重复执行的, 并依次生成最终变形场 (Figure 4), 该变形场编码了丰富的多尺度变形多层次背景信息。这允许模型在分层解码层上传播强上下文信息, 其中估计的变形场以粗到细的方式逐渐细化, 从而聚合高级上下文信息和低级详细特征。高级上下文信息使我们的模型具有处理大规模变形的能力, 而精细尺度特征使它能够建模详细的解剖结构信息。

### 3. 代码

有, 没看

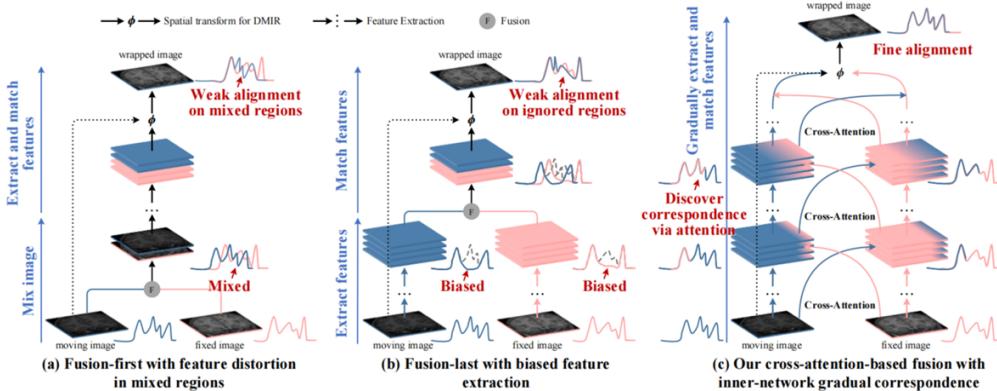
### 4. 问题

3D correlation layer 的实现细节没有搞清。

- [1] Lewis, K. M., Balakrishnan, G., Rost, N. S., Guttag, J., Dalca, A. V., 2018. Fast learning-based registration of sparse clinical images. arXiv: 1812.06932 .

## 17. XMorpher: Full Transformer for Deformable Medical Image Registration via Cross Attention (2022 MICCAI XMorpher)

### 1. 动机



- 虽然现有的深度网络在单幅图像特征表示方面具有较强的性能, 但这些单幅图像网络 (single image networks, SINs) 在DMIR中对一对图像的特征提取和匹配方面仍然存在局限性: (1) 混合区域特征失真的Fusion-first。如图Figure 1(a)所示, 一些DMIR方法将运动图像和固定图像进行融合, 模拟单幅图像输入条件, 并将融合后的图像发送到运动-固定特征的SIN中。但这些方法将特征提取和特征匹配过程混合在一起, 导致混合区域的特征失真和弱对齐, 从而使网络无法识别图像对之间的一对一对应关系。低效的特征表示能力最终导致了关键结构的缺失和较差的注册细节。 (2) 带有偏差特征提取的Fusion-last。如Figure 1(b)所示, 这些网络分别向双捷联惯导系统发送运动图像和固定图像, 最后融合来自不同网络的特征。但这些网络将特征提取和特征匹配过程绝对分离, 导致来自不同SINs的两个有偏差的特征最终匹配, 从而使

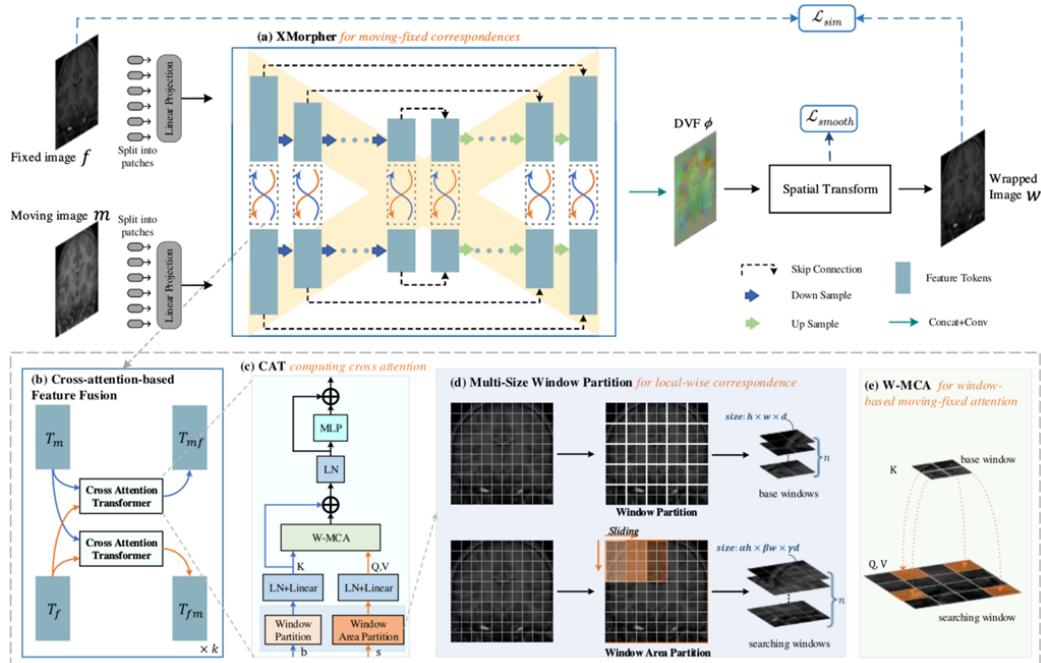
网络忽略了某些区域特征的不同层次（如多尺度）。特征表示的单一性限制了图像之间不同信息的对应关系，最终导致配准效果不佳。

- Transformer的注意机制因其出色的图像相关性捕捉能力而在配准中提供了潜在的应用，但现有的Transformer研究仅集中在单张图像上。在DMIR中，两个图像之间的移动-固定对应关系缺乏相关设计。用于DMIR的Transformer仍然采用与单幅图像任务相同的注意机制，只关注一张图像的相关性，而忽略了图像对之间的对应关系。动态图像与固定图像之间的对应关系捕获能力限制了变压器寻找有效的配准特征进行精细配准。

- 提出了一种用于双图像输入的新型Transformer——X-shape

Morpher (XMorpher)，将交叉注意引入到变压器结构中，实现高效、多层次的语义特征融合，有效地提高了配准性能。贡献如下：(1) 我们提出了一种新型的全Transformer主干网络。如Figure 1(c)所示，它包括两个并行的特征提取子网络，它们各自的特征以交叉注意的形式进行融合和匹配。通过递进交换网络，通过基于交叉注意的融合模块逐步融合匹配不同图像的特征，从而实现浮动和固定图像对应的有效特征表示，获得细粒度的多层次语义信息进行精细配准。(2) 提出了一种新的注意机制，Cross Attention Transformer (CAT) block，用于移动图像和固定图像的一对特征之间的充分通信。CAT块利用注意机制计算相互相关性，从而学习两张图像之间的对应关系，并促进特征在网络中自动匹配。(3) 基于DMIR的局部变换将特征匹配过程约束在窗口内，缩小了移动图像与固定图像之间的搜索范围，提高了计算效率，减少了大空间匹配时相似结构的干扰。这种基于窗口的特征通信极大地提高了配准的准确性和效率。

## 2. 方法



- XMorpher利用双U型网络分别提取运动图像和固定图像的特征（图(a)），两个网络通过特征融合模块进行通信，形成X型网络。两个并行网络遵循Unet的编码和解码部分的结构，但用CAT块代替卷积，在两个网络之间的注意特征融合模块中发挥重要作用（图(b)）。通过并行通信网络，我们的XMorpher垂直交换交叉图像信息，并在水平上不断细化特征。因此，最终输出的特征具有较强的表达运动图像和固定图像之间对应关系的能力。
- 来自并行子网络的对应特征 $T_m$ 和 $T_f$ 通过交换输入顺序，通过两个CAT块获得相互关注。然后两个输出与对方的注意力返回到原始pipeline，并为下一个更深入的通信做准备。为了获得足够的相互信息，总共有 $k$ 次通信。通过两个网络之间的注意特征融合模块，来自不同网络的具有不同语义信息的特征频繁交流，从而不断学习多层次的语义特征，最终进行精细配准。

- CAT块旨在通过注意机制计算输入特征b到特征s之间具有相应相关性的新特征 token (图 (b) )。b和s分别以不同的方式划分为基窗集 $S_{ba}$ 和搜索窗集 $S_{se}$ 两组窗口，用于下一个窗口的注意力计算。 $S_{ba}$ 和 $S_{se}$ 的大小n (窗口数量) 相同，但窗口大小不同。 $S_{ba}$ 中的每个基窗口通过线性层投影到查询集 $Q_{ba}$ ，每个搜索窗口通过线性层投影到 $K_{se}$ 和 $V_{se}$ 。然后，基于窗口的多头注意力 (W-MCA) 计算两个窗口之间的注意力，并将该注意力添加到基窗口中，使每个基窗口从搜索窗口中获得相应的加权信息。最后，将新的输出集发送到具有GELU非线性映射的2层MLP，以增强学习能力。在每个W-MCA和每个MLP模块之前应用一个LayerNorm (LN) 层。
- 针对变形图像配准主要关注体素局部位移，且移动图像与固定图像之间不存在大跨度对应关系的问题，提出了基于窗口的交叉注意机制，利用多尺寸窗口分区限制窗口内的注意计算。多大小窗口分区包括WP和WAP两种方法，将输入特征令牌b和s划分为不同大小的窗口，如图2 (d) 所示。基窗口集 $S_{ba}$ 的大小为 $n * h * w * d$ , n是窗口数量。为了获得相同数量的两个窗口集，WAP采用滑动窗口，并将步幅设置为基窗口大小， $S_{se}$ 的大小为 $n * \alpha h * \gamma w * \beta d$ 。
- Window-based Multi-head Cross Attention (W-MCA):

$$W - MCA(Q_{ba}, K_{se}, V_{se}) = softmax\left(\frac{Q_{ba}K_{se}^T}{\sqrt{d}}\right)V_{se}$$

### 3. 代码

有，没看

### 4. 问题

无

## 18. Learning Dual Transformer Network for Diffeomorphic Registration (2021 MICCAI)

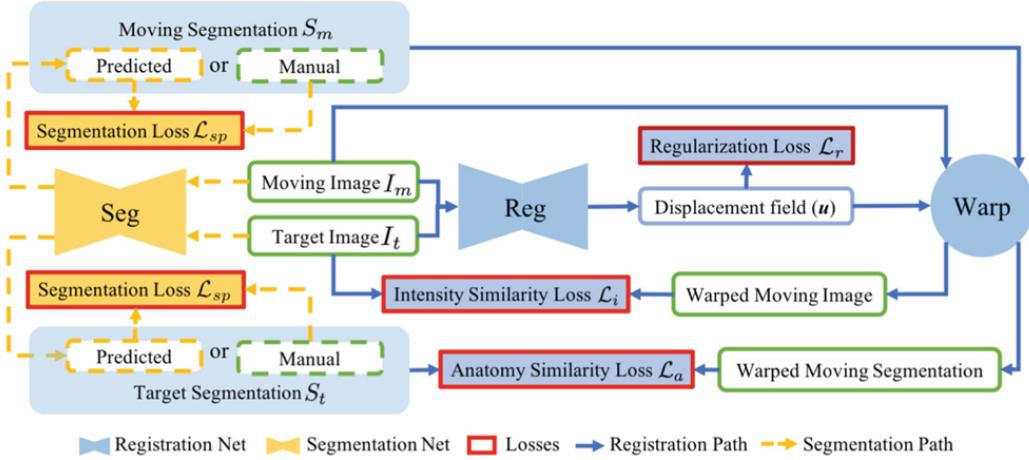
## 19. DeepAtlas: Joint Semi-supervised Learning of Image Registration and Segmentation (2019 MICCAI)

### 1. 动机

图像分割和配准可以相互促进，配准为分割提供了数据增强，分割为配准提供了额外的监督信息并用于评估配准结果。传统的联合分割和配准的方法对单个图像对进行操作而不是图像的总体，并且计算代价高。此外，获得 3D 医学图像的分割标签是困难的和劳动密集型的。因此，大部分 3D 图像数据都没有标签用于有监督学习。针对这样的情况，作者提出了 DeepAtlas，联合学习深度网络进行弱监督配准和半监督分割，贡献如下：

- 第一个提出了联合学习两个深度神经网络进行图像配准和分割。DeepAtlas 既可以联合训练，也可以单独训练和预测。
- DeepAtlas 只需要少量的人工分割标签，使用结构相似性损失来相互指导分割和配准。
- 在极端情况下，如果只有一个手动分割的图像可用，DeepAtlas 有助于 one-shot 分割，同时提高配准性能。

### 2. 方法



弱监督配准学习（蓝色实线部分）：图像相似性损失 $L_i$ ，正则化损失 $L_r$ 和解剖相似性损失（即分割标签的Dice损失） $L_a$ 的加权相加来训练配准网络。

半监督分割学习（黄色虚线部分）：有监督分割损失 $L_{sp}$ 和解剖相似性损失（通过配准网络扭曲的 $I_m$ 的分割标签和 $I_f$ 的分割标签的Dice损失） $L_a$ 的加权相加来训练分割网络，损失函数定义如下：

$$L_{seg} = \begin{cases} \lambda_a L_a(S_m \circ \Phi^{-1}, F_S(I_t)) + \lambda_{sp} L_{sp}(F_S(I_m), S_m), & \text{if } I_t \text{ is unlabeled;} \\ \lambda_a L_a(F_S(I_m) \circ \Phi^{-1}, S_t) + \lambda_{sp} L_{sp}(F_S(I_t), S_t), & \text{if } I_m \text{ is unlabeled;} \\ \lambda_a L_a(S_m \circ \Phi^{-1}, S_t) + \lambda_{sp} L_{sp}(F_S(I_m), S_m), & \text{if } I_t \text{ and } I_m \text{ are labeled;} \\ 0, & \text{if both } I_t \text{ and } I_m \text{ are unlabeled.} \end{cases} \quad (1)$$

当 $I_t$ 没有手动分割标签时， $L_a$ 相当于有监督分割损失，其中 $S_m \circ \Phi^{-1}$ 是噪声标签。当 $I_m$ 没有手动分割标签时，通过分割网络获得标签，然后扭曲这个标签并与 $S_t$ 计算Dice损失。当 $I_t$ 和 $I_m$ 都有标签时， $L_a$ 不监督分割网络，因为此时 $L_a$ 中没有 $F_s$ ，但会监督配准网络。当 $I_t$ 和 $I_m$ 都没有标签时，不训练分割网络。总的来说， $I_m$ 和 $I_t$ 谁没有手动分割标签，谁就通过分割网络来生成伪标签然后用于半监督训练，最少也要有一个手动分割标签。

训练时，交替训练两个网络中的一个，同时保持另一个固定。由于分割网络收敛更快，分割和配准网络交替训练步数为1:20。由于从零开始联合训练是很困难的，所以作者首先分别对单个分割、配准网络进行预训练。当真实标签数量极少时，比如只有一个，那么从零开始单独训练分割网络是很难的，所以作者最先使用无监督预训练好配准网络，然后再使用这个配准网络从头训练分割网络。直到分割网络能得到合理的结果后，才开始联合训练（交替训练）。

### 3. 总结

作者提出了DeepAtlas框架，用于仅使用少量标注图像的分割和配准网络的联合学习。当只给出一个真实分割标签时，作者的方法提供了one-shot分割学习，大大提高了配准效果。这表明，一个网络可以受益于对另一个网络提供的无标签数据的不完善监督。DeepAtlas为训练分割和配准网络时缺少真实分割标签提供了一个通用的解决方案。对于未来的工作，为分割和配准网络引入不确定性措施可能有助于缓解一个网络的不良预测对另一个网络的影响。研究通过层共享的分割和注册网络的多任务学习也将是有益的。这可能会进一步提高性能并减小模型尺寸。

### 4. 问题

在实验部分，作者考虑one-shot的情况（ $N = 1$ ， $N$ 表示手动分割标签的数量），设计了Semi-DeepAtlas（Semi-DA）：固定无监督（ $N = 0$ ）预训练好的配准模型，用于从零训练分割网络（ $N = 1$ ）。使用Semi-DA分割网络和无监督配准网络初始化DA模型。似乎 $N=1$ 的时候不足以训练好一个分割网络？没有理解它的one-shot分割过程。

# 20. A Cross-Stitch Architecture for Joint Registration and Segmentation in Adaptive Radiotherapy (2020 PMLR)

## 1. 动机

医学图像自动轮廓化的两种常用方法是图像分割和基于配准的轮廓传播。在自适应图像引导放射治疗的背景下，基于配准的方法具有使用患者解剖结构的先验知识的优势，并且能够准确deform低对比度结构，这些结构难以用附近高对比度结构进行识别。图像分割有其自身的优势，最显著的是能够准确地勾画出器官的轮廓，这些器官在两次访问之间的形状变化很大。

为了充分利用这两种方法的独特优势，提出了联合配准与分割 (JRS) 方法。在这项工作中，作者使用多任务学习领域的概念，通过在架构级合并这两个任务来进一步连接配准和分割，而不仅仅是通过损失函数。

## 2. 方法

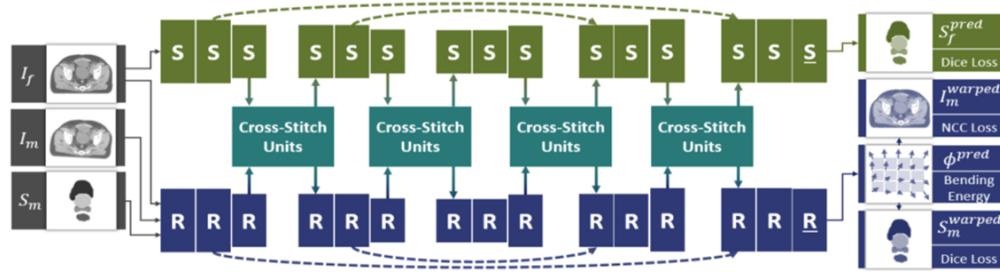


Figure 2: The inputs, architecture, outputs, and losses of the cross-stitch network.

作者提出的方法如上图所示。JRS输入固定图像 $I_f$ ，移动图像 $I_m$ ，和 $I_m$ 的分割标签 $S_m$ 。JRS输出 $S_f^{pred}$ ，并与真实标签 $S_f$ 计算Dice loss；输出形变场 $\Phi^{pred}$ ，并计算它的Bending Energy ( $\Phi^{pred}$ 的二阶导数，即Hessian矩阵) 作为正则化loss；用 $\Phi^{pred}$ 扭曲 $I_m$ 得到 $I_m^{warped}$ ，并与 $I_f$ 计算NCC loss；用 $\Phi^{pred}$ 扭曲 $S_m$ 得到 $S_m^{warped}$ ，并与 $S_f$ 计算NCC loss，这几部分loss加权相加作为总的loss。图中S表示分割层（一个或多个卷积组成的模块），R表示配准层。Cross-Stitch单元是这些层交换信息的模块[1]。Cross-Stitch的计算过程如下：给定分割网络S第l层的第k个卷积核得到的特征图 $X_S^{l,k}$ ，配准网络R第l层的第k个卷积核得到的特征图 $X_R^{l,k}$ ，和四个可学习的参数 $\alpha_{SS}^{l,k}, \alpha_{SR}^{l,k}, \alpha_{RS}^{l,k}$ 和 $\alpha_{RR}^{l,k}$ ，Cross-Stitch单元计算得到的特征图为：

$$\begin{bmatrix} \hat{X}_S^{l,k} \\ \hat{X}_R^{l,k} \end{bmatrix} = \begin{bmatrix} \alpha_{SS}^{l,k}, \alpha_{SR}^{l,k} \\ \alpha_{RS}^{l,k}, \alpha_{RR}^{l,k} \end{bmatrix} \begin{bmatrix} X_S^{l,k} \\ X_R^{l,k} \end{bmatrix} \quad (1)$$

Cross-Stitch的优点是能够学习在任务之间强烈共享特征映射，如果这是有益的。相反，如果特征映射对完全独立更好，网络可以学习单位矩阵来分离这些特征映射。这允许以一种灵活的方式在两个路径之间共享表示，在参数数量方面的成本可以忽略不计。

[1] Misra I, Shrivastava A, Gupta A, et al. Cross-stitch networks for multi-task learning[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). 2016: 3994-4003

## 3. 总结

在这项工作中，作者提出了从架构上连接图像配准和分割，以生成对自适应图像引导放疗至关重要的日常器官勾画。作者尝试了在三维全卷积神经网络中交织配准和分割的不同方法，发现用Cross-Stitch单元连接任务效果最好。通过Cross-Stitch单元，网络学习在其配准路径和分割路径之间交换信息。未来研究的一个有希望的方向是研究在联合网络中添加第三个任务，特别是放射治疗计划的生成。这可能使联合网络产生具有良好剂量学的特征。进一步的研究可能是针对不同患者群体和扫描仪的网络泛化。

#### 4. 问题

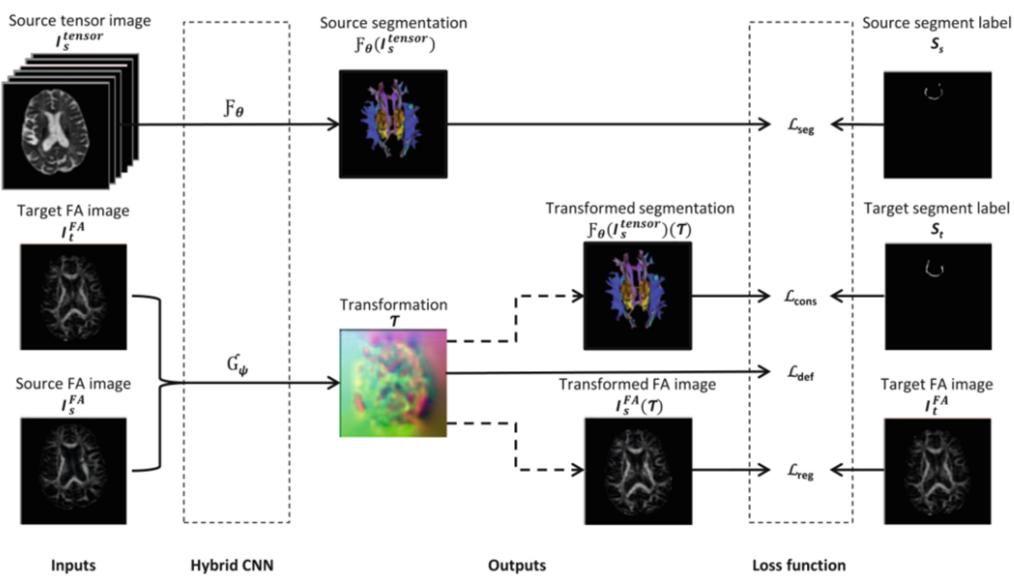
无

## 21. A Hybrid Deep Learning Framework for Integrated Segmentation and Registration: Evaluation on Longitudinal White Matter Tract Changes (2019 MICCAI)

#### 1. 动机

在纵向成像研究中，可以使用针对纵向数据定制的方法来提高分割的一致性。现有的解决方案通常涉及独立的配准和分割组件，这些组件在多级管道中按顺序或迭代地执行。利用可变形配准建立的空间对应关系，既可用来引入先验值在后续时间点进行分割，也可用来在公共空间中进行分割。作者在这里提出了一种新的混合卷积神经网络，它可以在单一过程中优化分割和配准。

#### 2. 方法



配准和分割共用一个网络。Hybrid CNN为U-Net。

#### 3. 总结

无

#### 4. 问题

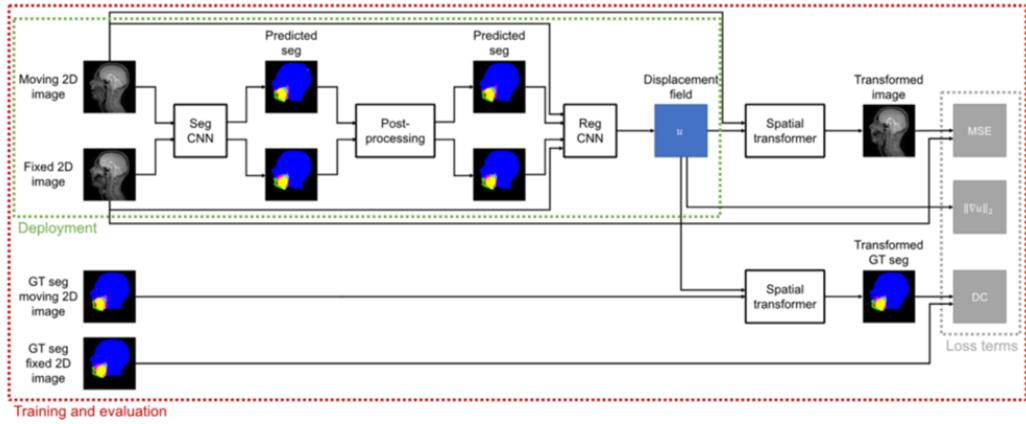
无

## 22. A segmentation-informed deep learning framework to register dynamic two-dimensional magnetic resonance images of the vocal tract during speech (2022 BSPC)

#### 1. 动机

动态磁共振(MR)成像可以在讲话过程中可视化发音器。在声道的二维MR图像中量化发音器运动的兴趣越来越大，以更好地理解语音产生，并可能为患者的管理决策提供信息。一些研究使用传统的变形配准方法来估计语音过程中声道的一系列动态2D MR图像中图像之间的位移场，然而，这些研究都没有评估或讨论配准方法是否捕捉到了舌头和软腭接触的变化。这项工作包括两个贡献。首先，它提出了基于分割的深度学习的可变形配准框架，以优化其在语音过程中估计声道动态2D MR图像之间的位移场。其次，这项工作首次使用了基于关节运动（腭咽闭合）的可量化和临床相关方面的度量来评估这些位移场的准确性。

#### 2. 方法



首先，图像对被用作分割网络的输入，分割网络估计图像中六个不同解剖特征的分割。其次，对分割进行后处理，以去除解剖学上不可能的区域（论文中没有细讲）。第三，将图像对和后处理分割用作配准网络的输入，该配准CNN估计位移场以使运动图像与固定图像对齐。第四，将运动图像和位移场作为空间变换器的输入，对运动图像进行变换。在训练和评估过程中，还使用空间变换器对运动图像的ground-truth (GT) 分割进行变换。损失函数包括移动图像和固定图像的相似度损失、形变场正则化损失和分割标签的Dice损失。

### 3. 总结

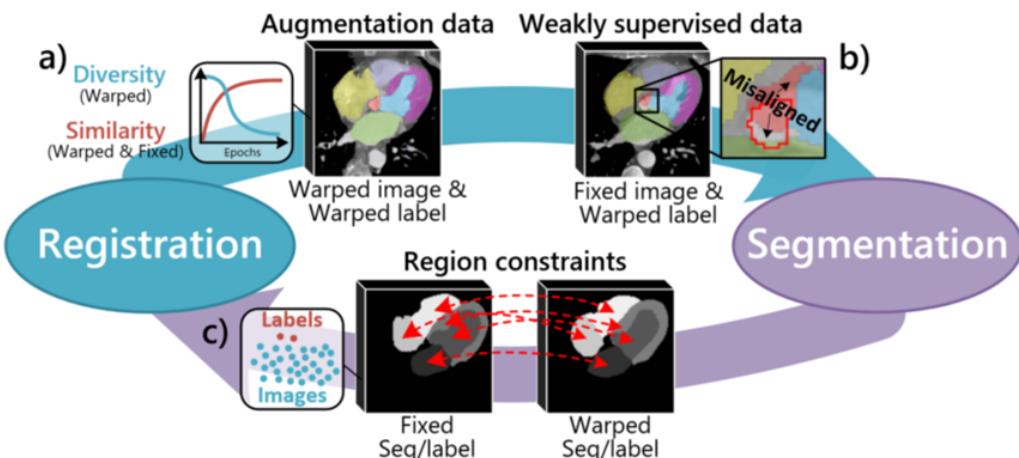
作者开发了一种用于估计语音过程中声道动态2D MR图像之间位移场的框架，并发现该框架比目前五种最先进的可变形配准方法和框架更准确地捕捉发音器运动的各个方面。该框架是朝着这类图像系列中关节运动的全自动量化的最终目标迈出的一步。此外，提出了一种基于发音器运动的临床相关和可量化方面的度量标准，并表明这对于评估语音动态MRI图像的注册框架是有用的。

### 4. 问题

无

## 23. Deep Complementary Joint Model for Complex Scene Registration and Few-Shot Segmentation on Medical Images (2020 ECCV)

### 1. 动机



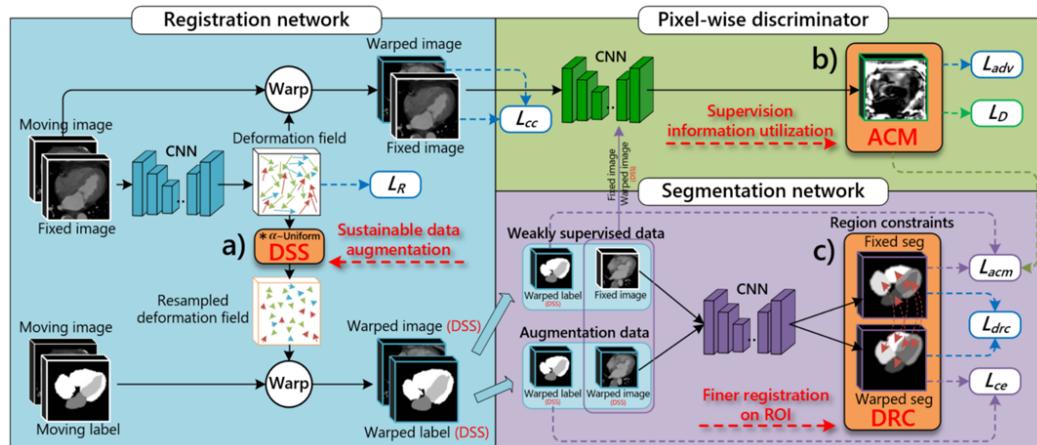
配准与分割任务具有很强的互补性，在复杂场景和few shot情况下可以相互促进。如上图所示，配准模型在训练过程中为分割模型提供了不同的增强数据（扭曲的图像和标签）或弱监督数据（固定图像和扭曲的标签），从而减少了标签的要求，增强了few shot情况下的分割泛化能力。分割模型对区域约束进行反馈，从而在复杂场景中更加关注感兴趣区域（ROI），实现更精细的配准。上图中的label应该是移动图像的label，此论文中固定图像应该没有ground true label。

然而，由于以下原因，这种互补拓扑的进一步利用受到阻碍：

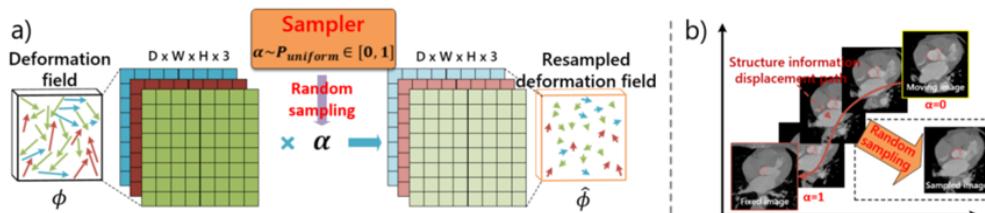
- Limitation 1：数据增强能力下降（上图 (a)）。配准模型在训练过程中，学习匹配真实情况的变形规则，生成不同的扭曲图像作为增强数据，提高分割泛化能力。然而，扭曲图像与固定图像之间的相似度增加并趋于稳定，随着相似度的稳定，扭曲图像的多样性逐渐减少。因此，在配准网络的后期训练阶段，在不同的epoch生成相同的扭曲图像，导致增强数据多样性降低。因此配准模型的数据增强能力下降，限制了分割的进一步增强。
- Limitation 2：弱监督数据中的错位区域（上图 (b)）。弱监督数据扩大了标记数据集，并为分割模型提供了额外的监督信息。但是，这些数据中较大的错位区域会产生不正确的优化目标，如果直接使用会扰乱训练过程，导致严重的误分割。
- Limitation 3：缺乏基于标签的区域约束（上图 (c)）。然而，在few-shot环境下，由于标签较少，缺乏基于标签的区域约束。因此在复杂场景下，配准模型会进行粗糙优化，复杂的背景会限制在ROI上的配准性能。

针对这三个问题，作者提出了三个解决方案（第2节叙述）。本文提出了一种深度互补联合模型（Deep Complementary Joint Model, DeepRS），该模型最小化复杂场景下的背景干扰，以实现对ROI的更精细配准，并大大降低了少镜头情况下分割的标签要求，以获得更高的泛化能力。

## 2. 方法



Solution 1: Deep Structure Sampling (DSS) for Sustainable Data Augmentation.



DSS块通过在变形场中嵌入随机扰动因子，持续生成不同的增强数据，以增加扭曲图像和标签的不确定性。配准过程是图像结构信息的位移，而形变程度的扰动实现了对该位移路径上的信息采样。因此，DSS块带来两个优势：1) 可持续的数据增强。通过扰动因子控制配准网络的变形程度，保证配准网络能够持续生成多样化的增强数据。2) 真实分布。从位移路径中提取结构信息，得到的增强数据比其他人工增强方法更符合实际分布。上图 a) 对形变场 $\phi$ 乘以从均匀分布中采样得到的扰动因子 $\alpha$ ，得到采样后的形变场 $\hat{\phi}$ 。因此，即使配准网络已经融合，被它扭曲的图像和变形的标签仍然会有很大的多样性。从图 b) 可以看出，随着 $\alpha$ 的增大，由于其结构信息接近于固定图像，扭曲图像逐渐接近于固定图像。

Solution 2: Alignment Confidence Map (ACM) for Supervision Information Utilization.

基于Patch-GAN的像素级鉴别器学习扭曲图像和固定图像之间的相似性，并输出突出显示对齐区域的对齐置信度图。因此，在计算弱监督损失函数时，通过这些对齐置信度图可以抑制不对齐的区域，并利用对齐区域中的监督信息进行更高的分割泛化，如式 (1) 所示

$$L_{acm} = -D(W(x_m, \hat{\phi})) W(y_m, \hat{\phi}) \log S(x_f) \quad (1)$$

其中， $x_m, y_m, x_f$ 和 $\hat{\phi}$ 分别表示移动图像、移动图像的标签、固定图像和采样的形变场。 $D(\cdot, \cdot)$ 表示计算两个图像直接的相似度， $W(\cdot, \cdot)$ 表示扭曲操作。 $D(\cdot, \cdot)$ 计算得到的应该还是一个矩阵，表示对应位置的像素的相似度。 $S(x_f)$ 应该是分割网络输出的软分类标签。 $W(y_m, \hat{\phi})$ 和 $S(x_f)$ 重合（对齐）的区域权重大，对应的相似度损失就越大，误差区域的损失值将得到较低的权重，从而抑制干扰。由于 $y_m$ 是0,1二值化的，不重合的部分权重应该是为0。这个公式中似乎没有对齐置信度图，即图中的绿色箭头。而且，Limitation 2说的应该是固定图像和扭曲得到的标签的不对齐问题，但这里解决的是输入到配准网络的扭曲标签和分割后的标签不对齐的问题？

Solution 3: Deep-Based Region Constraint (DRC) for Finer Registration on ROIs.

DRC策略通过来自分割网络的固定和扭曲分割掩码之间的约束（公式(2)）来引导注意力在ROI上进行更精细的配准。该深度区域约束以变形图像和固定图像中对应区域的对齐为优化目标，实现了1) 在少镜头情况下释放基于标签的区域约束的标签要求，2) 独立优化不同区域以避免相互之间的不对齐，3) 在ROI上额外关注区域以实现更精细的配准。

$$L_{drc} = -(S(W(x_m, \hat{\phi})) - S(x_f))^2 \quad (2)$$

即分别在分割网络中输入扭曲的固定图像和固定图像，输出两个分割图像，计算它们的MSE。每个ROI在不同通道中计算，得到独立的精细优化，**而任务不相关区域在一个后台通道中计算（不理解）**。因此，ROI上的精细配准是可用的，并避免了区域间的错误配准。

DeepRS模型中的配准网络、分割网络和像素级鉴别器通过不同的损失函数组合进行训练，以协调训练过程，实现相互改进。配准网络的损失函数为：

$$L_{reg} = \lambda_{adv} L_{adv} + \lambda_{drc} L_{drc} + \lambda_{cc} L_{cc} + \lambda_R L_R \quad (3)$$

其中，像素级鉴别器的对抗损失 $L_{adv}$ 提供了扭曲图像和固定图像之间的相似度度量，分割网络DRC的损失引起了对ROI的配准关注， $L_{cc}$ 表示局部互相关损失， $L_R$ 表示形变场的正则化损失。分割网络的损失函数 $L_{seg}$ 由两部分组成：

$$L_{seg} = \lambda_{acm} L_{acm} + \lambda_{ce} L_{ce} \quad (4)$$

ACM损失 $L_{acm}$ 将弱监督数据加入到训练中，以获得更高的分割泛化能力，扭曲标签和扭曲图像通过分割网络得到的标签之间的交叉熵损失 $L_{ce}$ 来保持正确的优化目标。鉴别器由参考图像 $x_r$ 和固定图像 $x_f$ 组成的配准图像对作为阳性情况，由扭曲图像 $x_w$ 和固定图像 $x_f$ 组成的图像对作为阴性情况。参考图像 $x_r$ 是运动图像 $x_m$ 和固定图像 $x_f$ 的融合， $x_r = \beta * x_m + (1 - \beta) * x_f$ 。鉴别器的损失除了 $L_{adv}$ 外，还有鉴别真假固定图像的损失 $L_D$

$$L_D = -\log(D(x_r, x_f)) - \log(1 - D(x_w, x_f)) \quad (5)$$

### 3. 总结

本文提出了一种用于复杂场景配准和少镜头分割的深度互补联合模型（DeepRS）。本文提出的DSS块通过扰动因子随机调整变形场，从而提高了扭曲图像和标签的活性，实现了可持续的数据增强能力；提出的ACM方法通过像素级鉴别器的对齐置信度映射有效地利用弱监督数据中的监督信息，带来更高的分割泛化；提出的DRC策略从分割模型中构建了扭曲和固定图像之间的无标签损失，从而在ROI上实现更精细的配准。本文的工作大大降低了对大型标记数据集的要求，并提供了精细的优化目标，从而提高了配准和分割精度，大大节省了成本。特别是，我们的DeepRS模型在一些标记困难、场景复杂或数据集小的情况下具有很大的潜力。

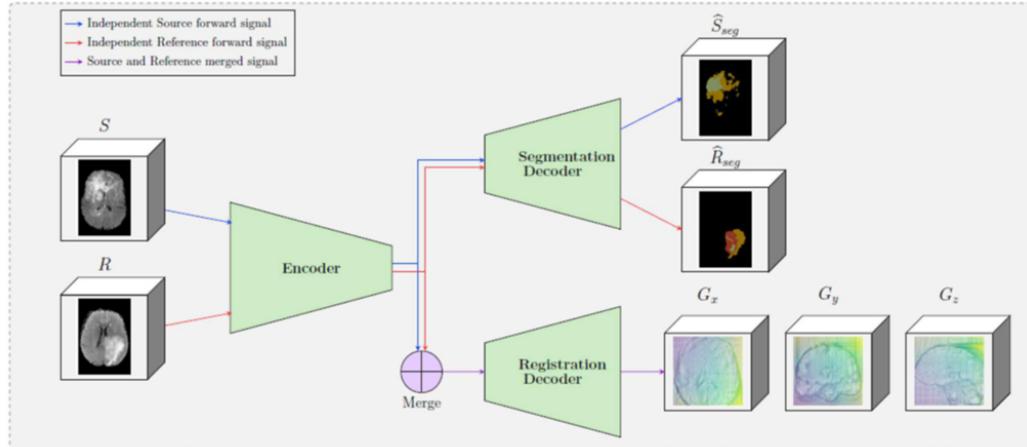
### 4. 问题

第2节加粗部分。

## 24. Deep Learning-Based Concurrent Brain Registration and Tumor Segmentation (2019 Frontiers in Computational Neuroscience)

### 1. 方法

本文中，我们提出了一种基于双重深度学习的架构，同时解决配准和肿瘤分割问题，放松了预测肿瘤区域内的配准约束，同时提供位移场和分割图像。



共享编码器，分离解码器结构。

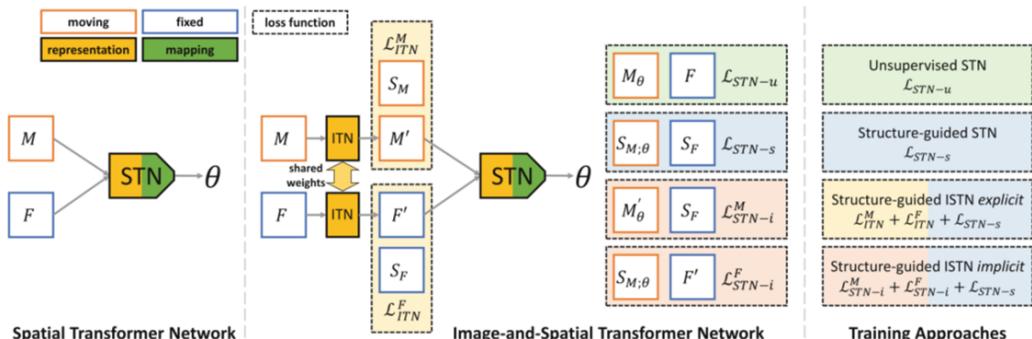
## 25. Image-and-Spatial Transformer Networks for Structure-Guided Image Registration (2019 MICCAI)

### 1. 动机

作者认为到目前为止，基于神经网络的图像配准并没有充分利用深度表示学习。同时观察到，无论监督方法还是无监督方法都没有利用神经网络的两个关键优势，即 1) 学习为下游任务优化的新表示的能力，以及 2) 在训练期间incorporate额外的信息并从中受益的能力，但是这些信息在测试时不可用或很难获得。这种额外的监督（如分割标签和landmark等）可以帮助在测试时以不同于单独使用图像强度的方式指导配准。例如，配准可能专注于特定的Structures-of-Interest (SoI)。然而，目前的方法不能保留或显式提取这些额外的信息，因此不能在测试时进一步使用。

为了克服这些限制，并充分利用神经网络学习表示的关键能力，引入了图像和空间转换网络 (ISTN)，其中添加了专用的图像转换网络 (ITN) 到空间转换网络 (STN) 的头部，旨在提取和保留有关SoI的信息。ITN产生一种新的图像表示，该图像表示以端到端方式学习，并针对下游配准任务进行优化。这不仅允许我们在测试时预测良好的初始转换，而且允许使用完全相同的模型进行精确的特定于测试的迭代细化，从而实现结构引导配准。

### 2. 方法



STN是大多数基于DL的图像配准网络的构建块。STN有两个主要组成部分：使用卷积层学习输入的新表示的特征提取部分，以及将这些表示映射到转换参数的第二部分。然而，STN可以学习的表示形式并不是公开的，而是在推理过程中保持隐藏（可能是指无法直观理解到这些表示的所代表的含义）。作者通过引入专用的图像转换网络，重新设计了基于图像配准的神经网络转换模块的基本构建模块。

作者将ITN定义为卷积神经网络，将输入图像映射到相同大小和维度的输出图像。ITN的作用是显式地公开学习的图像表示，这对于STN解决的下游配准任务是最优的。这种新架构提出了许多训练方法，特别是当关于Sol的额外信息可用时，例如图像分割或landmark。