Fully Convolutional Network for Consistent Voxel-Wise Correspondence

Yungeng Zhang,[1] Yuru Pei,[1*] Yuke Guo,[2] Gengyu Ma,[3] Tianmin Xu,[4] Hongbin Zha[1]
[1]Key Laboratory of Machine Perception (MOE), Department of Machine Intelligence, Peking University, Beijing, China
[2] Luoyang Institute of Science and Technology, Luoyang, China
[3] uSens Inc., San Jose, USA
[4] School of Stomatology, Peking University, Beijing, China
{zhangyungeng, yrpei}@pku.edu.cn, guoyuke02@aliyun.com,
{magengyu, tmxuortho}@gmail.com, zha@cis.pku.edu.cn

Abstract

In this paper, we propose a fully convolutional network-based dense map from voxels to invertible pair of displacement vector fields regarding a template grid for the consistent voxel-wise correspondence. We parameterize the volumetric mapping using a convolutional network and train it in an unsupervised way by leveraging the spatial transformer to minimize the gap between the warped volumetric image and the template grid. Instead of learning the unidirectional map, we learn the nonlinear mapping functions for both forward and backward transformations. We introduce the combinational inverse constraints for the volumetric one-to-one maps, where the pairwise and triple constraints are utilized to learn the cycle-consistent correspondence maps between volumes. Experiments on both synthetic and clinically captured volumetric cone-beam CT (CBCT) images show that the proposed framework is effective and competitive against state-of-the-art deformable registration techniques.

1 Introduction

To find the dense voxel-wise correspondence of a volume pair is an essential task of a variety of applications in medical images analysis (Sotiras, Davatzikos, and Paragios 2013), such as statistical shape analysis (Lombaert, Arcaro, and Ayache 2015) and the label propagation of predefined landmarks and segmentation (Kanavati et al. 2017). The dense correspondence can be used in the studies of the pre- and post-treatment assessments to find the structure progression due to longitudinal operations and growths, especially for adolescent patients.

The traditional deformable registration techniques rely on online non-linear iterative optimization to minimize the voxel-wise appearance difference, together with a regularization term to maintain the smoothness of the displacement vector fields (DVFs). Concerning the large set of parameters to be solved in the volumetric image registration, the optimization solving is computationally intensive and prone to be stunned in a local minimum (Pluim, Maintz, and Viergever 2003).

Recent works on the deformable registration (Rohé et al. 2017; Yang et al. 2017; Sokooti et al. 2017; Krebs et al. 2017; Balakrishnan et al. 2018; Dalca et al. 2018) have shown the merits of the CNN-based regression for dense correspondence. In the supervised deep learning framework, the ground-truth corresponding landmarks and DVFs are required for the training (Rohé et al. 2017; Yang et al. 2017; Sokooti et al. 2017; Krebs et al. 2017). Since the manual labeling of volumetric images is more laborious than the ordinary 2D images and prone to the practitioners' experiences, the learning suffers from the limited training data. The unsupervised frameworks follow the spatial transformer network (STN) to minimize the voxel-wise appearance differences (Balakrishnan et al. 2018; Dalca et al. 2016; 2018). The existing systems solve the patch-wise registration to relieve the memory burden (Dalca et al. 2016), or to solve the unidirectional maps to the template or the atlas (Balakrishnan et al. 2018; Dalca et al. 2016). In order to get the invertible one-to-one map, the additional diffeomorphic integration layer is needed to obtain the final registration field from the estimated velocity field (Dalca et al. 2018; Krebs et al. 2019).

There are several papers exploring the usage of cycle consistency to improve the maps computed between pairs of 3D shapes or 2D images. Huang et al. present a time-consuming optimization approach to compute a set of new maps of 3D shapes aligned with initial maps considering cycle consistency (Huang et al. 2012). Zhou et al. propose to utilize 4-cycle consistency as a supervisory signal to address correspondence between 2D images, leveraging additional 3D CAD models (Zhou et al. 2016). In this paper, we aim at finding cycle-consistent voxel-wise correspondence of a volume corpus and introduce a fully convolutional network to obtain both the forward and backward parameterized transformation functions between an input volume and a shared template. Following the STN (Jaderberg et al. 2015), we minimize the gap between the warped volumetric image and the target and learn the network in an unsupervised way. We introduce a combinational inverse constraint to enforce the invertibility of forward and backward transformation pair regarding the template for the one-to-one mapping. Further, we utilize the arbitrary pair and triplet volumes in

[*]Corresponding Author

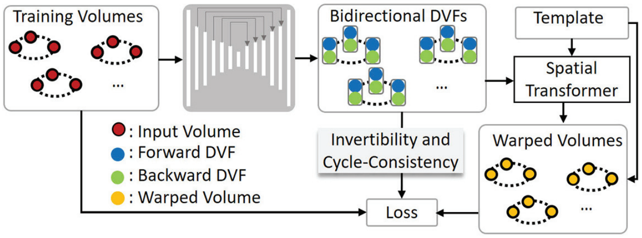

Figure 1: The framework of the proposed network for consistent voxel-wise correspondence.

the datasets to enforce the cycle-consistency of the transformations, where the transformation concatenation along a closed cycle with two or three volumes are required to be an identical transformation. For the registration of fine-grained anatomies, we introduce an optional bidirectional structural constraint. The online deformable registration between volumes is obtained by a simple evaluation of the learned mapping function instead of the expensive iterative optimization. The novelty of this paper is:

- We present an unsupervised learning-based framework for dense consistent voxel-wise correspondence;
- We propose CNN-based parameterized functions for both the forward and backward transformations and enforce the invertible and cycle-consistent registrations of the volume corpus;
- Our method enables an efficient online dense voxel-wise correspondence estimation for clinically captured volumetric images.

2 Related Work

3D deformable medical image registration has been addressed extensively in past decades (Sotiras, Davatzikos, and Paragios 2013). The conventional techniques employ several metrics, such as the mean squared distance, the normalized cross-correlation (CC) (Avants, Epstein, and Gee 2008), and the mutual information (Maes and others 1997) to minimize the appearance differences by a large scale nonconvex optimization. Several physical and interpolation-based models have been used in the deformable registration, including the elastic body model (Bajcsy and Kovacic 1989), the demons (Thirion 2011), the flow of diffeomorphisms (Beg et al. 2005), and the B-spline-based free form deformations (Rueckert et al. 1999). The diffeomorphism transform realizes the invertible and smooth one-to-one map and avoids the structure folding with topology preservation, which is a usually desirable property of the anatomical image registration. The popular formulations, including the large diffeomorphic distance metric mapping (LDDMM) (Beg et al. 2005), the symmetric normalization (SyN) (Avants, Epstein, and Gee 2008), and the diffeomorphic demons (Vercauteren et al. 2009) have been used in anatomy studies. The optimization-based deformable registration is known to be time-consuming considering the large set of involved voxels and the parameters to be solved. The subsampling (Roshni, Fessler, and Boklye 2009) and statistical deformation model (Ashburner and Friston 2000) are used to reduce the problem space, which also relies on the iterative optimization for parameter solving.

The learning-based deformable registration techniques avoid online optimization by finding the regression functions from the input images to the registration parameters. The function evaluation by the support vector regression (Minjeong et al. 2012), the random forests (Wei et al. 2017), and the recent deep neural networks (de Vos et al. 2017; Rohé et al. 2017; Yang et al. 2017; Sokooti et al. 2017; Krebs et al. 2017; Balakrishnan et al. 2018; Dalca et al. 2018) have a magnitude smaller time complexity than conventional optimization-based methods. The unsupervised frameworks following the STN (Jaderberg et al. 2015) avoid data annotation in the training process. In the CNN-based probabilistic generative model for the diffeomorphic registration (Dalca et al. 2018; Krebs et al. 2019), the additional integration layer is used to obtain the final registration field. There exist studies addressing the inverse constraints, which avoid the intermediate flow estimation by inferring the bidirectional DVFs directly (Christensen and Johnson 2001; Leow et al. 2005; He and Christensen 2003). However, they rely on the hand-crafted features. The inverse constraints are just imposed on the image pairs without considering the consistent correspondence in the image corpus. For instance, the image warped by the transformation concatenation along an arbitrary and closed cycle of volumes should be identical to itself. In this paper, we explicitly infer dense and cycle-consistent correspondence of volumetric images without need of intermediate velocity flow estimation.

3 Method

We follow the template deformation paradigm and introduce the bidirectional network-based mapping functions for the registration between volume image V and a prototypical or template volume T. Let $V, T \in \mathbb{R}^3$ be single-channel grayscale volumes in a 3D spatial domain. We utilize the CNN to model the mapping function $h_{\Theta,T} : V \to [\phi_f, \phi_b]$. Function h_Θ bridges the input image V with the bidirectional transformations, $\phi_f : V \to T$ and $\phi_b : T \to V$. Θ denotes the learnable parameters of function h, i.e., the kernel weights of the convolutional network. Instead of inferring unidirectional mapping between volumes V and T, the output of our system is a pair of DVFs $[\phi_f, \phi_b] \in \mathbb{R}^6$. For voxel $x \in \mathbb{R}^3$, $\phi(x)$ returns the location of the voxel's counterparts in the target volume. The volume pair $(V \circ \phi_f, T)$ and $(V, T \circ \phi_b)$ are expected to bear similar anatomical appearances.

Fig. 1 shows an overview of the proposed method. In the training phase, the system takes one volume $V \in \mathcal{V}$ as input, and estimate $[\phi_f, \phi_b]$ using the parameter Θ. The spatial transformation layer realizes the image warping and results to $V \circ \phi_f(x)$ and $T \circ \phi_b(x)$. We evaluate the volume similarity in the forward and backward directions to find the optimal parameters Θ. Moreover, we introduce a combinational inverse constraint to enforce a pair-wise invertible DVFs and the cycle-consistent registrations in a corpus. Given a train-

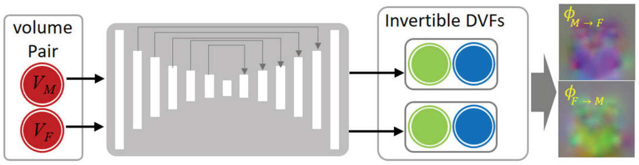

Figure 2: The flowchart of online registration and dense correspondence between a volume pair (V_M, V_F).

ing dataset $\mathcal{V} = \{V_i | i = 0, \ldots, N\}$ with a distribution of $\mathcal{D}_{\mathcal{V}}$, we minimize the expected loss functions $\mathcal{L}$ for Θ:

$$\Theta = \arg\min_{\Theta} \mathbb{E}_{V \in \mathcal{D}_{\mathcal{V}}} \mathcal{L}(V, h_{\Theta,T}(V)). \tag{1}$$

In the testing phase (see Fig. 2), given an arbitrary input pair of the moving and fixed volumes (V_M, V_F), the network returns the forward and backward DVFs $[\phi_f^M, \phi_b^M]$ and $[\phi_f^F, \phi_b^F]$ of V_M and V_F respectively, then we obtain the invertible and consistent DVFs $\phi_{M\to F} = \phi_f^M \circ \phi_b^F$ and $\phi_{F\to M} = \phi_f^F \circ \phi_b^M$.

3.1 DVF Inference

We employ the fully convolutional network to parameterize function $h_{\Theta,T}$. Similar to the 3D U-net (Çiçek et al. 2016), we use the symmetric encoder-decoder structure with long residual connections to infer the 6-channel registration pair $[\phi_f, \phi_b]$ from the 1-channel volume V. In our system, the encoder has six $3 \times 3 \times 3$ convolutional layers of stride 1 followed by the instance normalization and Leaky ReLU. A $2 \times 2 \times 2$ pooling layer follows each convolutional layer. The decoder module also has six 3D deconvolutional layers with Relu and instance normalization. We use a convolution with a fractional stride of 1/2 for upsampling of feature volumes in the decoder for the DVF with the same resolution of the input. The skip connections between encoder and decoder facilitate feature propagation and fast convergence.

Spatial Transformation. We follow the unsupervised STN (Jaderberg et al. 2015) to obtain the warped volumes using the estimated bidirectional DVFs. In order to estimate the intensity value on the regular grid of the warped volume, we linearly interpolate the neighboring voxels in the surrounding cube. In the volumetric image interpolation of voxel x for the bidirectional warping, we consider the neighboring voxels in $\mathcal{N}(\phi_\kappa(x)), \kappa \in \{f, b\}$.

$$V \circ \phi_\kappa(x) = \sum_{x' \in \mathcal{N}(\phi_\kappa(x))} V(x') \Pi_{i=0}^{2}(1 - |x_i - x_i'|). \tag{2}$$

The operator is differentiable almost everywhere, which enables the error back-propagation in the optimization.

3.2 Consistent Voxel-wise Correspondence

The goal is to estimate the consistent voxel-wise correspondence by the CNN-based bidirectional DVF prediction framework. We begin with introducing definitions of the consistent correspondence of a volume corpus.

Definition 1. *(Invertibility) A transformation ϕ between volumetric image V_i and V_j satisfies the invertibility property, if both $\phi_{i\to j} \circ \phi_{j\to i}$ and $\phi_{j\to i} \circ \phi_{i\to j}$ are identical transformations. $\phi_{i\to j} : V_i \to V_j$ and $\phi_{j\to i} : V_j \to V_i$ denote the forward and backward transformations between V_i and V_j.*

Definition 2. *(Cycle Consistency) A set of transformations $\Phi = \{\phi_{p\to q} | p, q = 0, \ldots, N; p \neq q\}$ in a volume corpus are consistent if for an arbitrary circle $V_0 - V_1 - \cdots - V_k - V_0$ composed of $k+1$ volumes, the concatenation of the DVFs,*

$$\phi_{0\to 1} \circ \phi_{1\to 2} \cdots \circ \phi_{k\to 0}, \tag{3}$$

is an identical transformation.

Proposition 1. *If there exist a set of invertible transformations $\Phi = \{\phi_{\{f,b\},i} | i = 0, \ldots, N\}$ from a volume corpus $\mathcal{V}$ to a common latent volume, the set Φ defines a group of consistent registrations of $\mathcal{V}$, which satisfy the cycle consistency property.*

Proof. Given one arbitrary cycle $V_0 - V_1 - \cdots - V_k - V_0$ composed of $k+1$ volumes, the concatenation of DVFs along the cycle is defined as: $\phi_{0\to 1} \circ \phi_{1\to 2} \cdots \circ \phi_{k\to 0}$.

If there exist a set of invertible transformations $\Phi = \{\phi_{\{f,b\},i} | i = 0, \ldots, k\}$ from volume $\{V_i | i = 0, \ldots, k\}$ to a template T, the DVF $\phi_{p\to q}$ between image V_p and V_q is computed as $\phi_{p\to q} = \phi_{f,p} \circ \phi_{b,q}$. Then, the warping of V_0 by the concatenation of the DVFs is as follows:

$$\begin{aligned} & V_0 \circ \phi_{0\to 1} \circ \phi_{1\to 2} \cdots \circ \phi_{k\to 0} \\ &= V_0 \circ (\phi_{f,0} \circ \phi_{b,1}) \ldots (\phi_{f,k} \circ \phi_{b,0}) \\ &= V_0 \circ \phi_{f,0} \circ \phi_{b,0} = V_0. \end{aligned}$$

□

We introduce a combinational inverse constraint to learn the end-to-end mapping functions for the DVFs that satisfy the invertibility and cycle-consistent properties. Considering the Proposition 1, the bidirectional transformations between image V_i in the training dataset and the template grid T are required to be invertible. We further expect the transformation concatenation along an arbitrary circle in the training dataset be an identical one. We define the inverse loss function as follows:

$$\begin{aligned} L_{inv} = & \sum_{i=1}^{m} \{\|\phi_{f,i} \circ \phi_{b,i}\|_F^2 + \|\phi_{b,i} \circ \phi_{f,i}\|_F^2\} \\ & + \sum_{k=1}^{K} \sum_{\substack{c \in \mathcal{C}, \\ |c|=k}} \|\Pi_{j=0}^{k-1} \phi_{j\to mod(j+1,k)}\|_F^2. \end{aligned} \tag{4}$$

The first term is used to enforce the bidirectional transformations $[\phi_f, \phi_b]$ between the training corpus with the shared template to be invertible. m denotes the volume number in the mini-batch, and is set to 3 in our experiments. $\|\cdot\|_F$ denotes the Frobenius norm.

In the second term, we require the DVF concatenation $\Pi_{j=0}^{k-1} \phi_{j\to mod(j+1,k)}$ along a closed circle c with k volume images $\{V_i | i = 0, \ldots, k-1\}$ to be an identical transformation. For a training dataset with N volume images, the number of the closed circle set $\mathcal{C}$ of length more than 2 is $\sum_{k=2}^{N} \frac{n!}{(n-k)!}$. For simplicity, we only consider cycles with pair or triplet volumes, and K is set at 3 in our experiments. The second term enforces the invertibility of DVFs of an arbitrary image pair and the cycle-consistency of a triplet. Note that in the testing stage, our system takes an arbitrary moving and fixed volume pair as an input and returns the bidirectional DVFs. The case of $k = 2$ in the second term of Eq. 4 enforces the consistent voxel-wise correspondence between the arbitrary volume pair. We compute the DVF concatenation in L_{inv} using a coordinate volume Q, which is transformed by the DVFs to Q'. The DVF concatenation is simply defined as $Q' - Q$.

3.3 Loss Function

As shown in Fig. 1, we learn the parameterized mapping functions from the volumetric image to the bi-directional DVFs for the consistent voxel-wise correspondence. By minimizing the loss function, we try to find the optimal network parameters Θ.

$$\mathcal{L}(\Theta) = \alpha_{sim} L_{sim} + \alpha_{reg} L_{reg} + \alpha_{inv} L_{inv}. \tag{5}$$

In our system, the CNN-based functions parameterize the mapping between a volume in the training dataset and a shared template. Given the estimated forward and backward DVFs using network parameters Θ, the similarity term L_{sim} measures the gap between the warped volumes with the target in both the forward and backward directions. Here we use the cross-correlation similarity metric to measure the volumetric image difference.

$$L_{sim} = \sum_{i}^{m} \{\|V_i \circ \phi_{f,i} - T\|^2 + \|V_i - T \circ \phi_{b,i}\|^2\}. \tag{6}$$

The regularization term L_{reg} enforces the smoothness of both the forward and backward DVFs. We apply the diffusion regularizer on the spatial gradients of the forward and the backward DVFs.

$$L_{reg} = \sum_{i=1}^{m} \{\| \bigtriangledown \phi_{f,i}\|_F^2 + \| \bigtriangledown \phi_{b,i}\|_F^2\}. \tag{7}$$

The spatial gradients are approximated using the numerical differences in the x-, y-, and z-directions.

The constant coefficients α are used to balance the terms regarding the similarity, the regularization, and the inverse constraints. In our system, we set the parameters as follows: $\alpha_{sim} = 1$, $\alpha_{reg} = 50$, $\alpha_{inv} = 50$.

3.4 Optional Structure-aware Constraint

In order to improve the registration accuracy of fine-grained structures, such as the anterior cranial base (ACB) in the craniofacial CBCT images, we introduce an optional structure-aware loss L_{str} to penalize the inconsistency in the structure of interests (SOI) set.

$$\begin{aligned} L_{str} = & \sum_{i=1}^{m} \{\|M \star (V_i \circ \phi_{f,i} - T)\|^2 \\ & + \|(M \circ \phi_{b,i}) \star (V_i - T \circ \phi_{b,i})\|^2\}. \end{aligned} \tag{8}$$

The operator $\star$ denotes the per-element matrix multiplication. M denotes the mask of SOIs with entry set at 1 for voxels inside the SOI and 0 otherwise. We only define the SOI mask on the template image T, and do not require any SOI annotation in the training and testing data. The first part of L_{str} measures the inconsistency of SOI between the warped V_i and the template. Since the SOI is defined on the shared template, in the second part regarding the backward registration, we warp the mask using the backward DVFs as $M \circ \phi_b$ to the space of input volume V_i.

3.5 Training Details

Instead of training the convolutional network from scratch, we pre-train the network using a set of synthetic volumes. We generate the synthetic volumes using random B-spline-based deformation of the template. The resulted forward and backward DVFs are used to train the network. Note that the network initialized by the synthetic volumes is limited to handle the bidirectional DVF prediction as stated in the experimental section because the synthetic dataset could not cover the shape and appearance variations of the volume corpus.

We train the network using the ADAM optimizer with a learning rate of 1e-4 and momentums of 0.5 and 0.999. The mini-batch contains three volumes. The framework is implemented using the open-source PyTorch implementation of convolutional neural networks on an NVIDIA GTX TITAN X GPU. The training takes 66 hours of 300 epochs. The testing of the registration between an arbitrary volume pair takes 0.17s.

4 Experiments

Dataset. We validate the proposed method on craniofacial CBCT images. The training dataset consists of 400 clinically captured CBCT images from orthodontic patients, including both the pre- and post-treatment volumes. The volume image resolution is $128 \times 128 \times 128$. The size of the isotropic voxel is $1.5 \times 1.5 \times 1.5mm^3$. For testing, we collect a toy dataset with 20 synthetic images and a real dataset with 20 clinically captured images. The voxel values are normalized to $[-1, 1]$. It is not easy to get the ground-truth DVFs, so we generate a toy dataset with the ground-truth DVFs using synthetic data, where the template volume is deformed using arbitrary B-spline-based deformations without structure folding. The control grid is set at $7 \times 7 \times 7$. In our system, we define the template as the statistical shape average of the training dataset (visualized in Fig. 4) to avoid the bias in the DVF estimation.

Metrics. The registration accuracy is evaluated by the mean squared distance (MSD) of the predicted forward and backward DVFs with the ground-truth in the toy dataset. Except for the measurement of MSD, all testing experiments are conducted on the clinically captured images. For the quantitative assessment of the consistent voxel-wise correspondence on the clinically captured images, we use the Dice similarity coefficient (DSC) in the label propagation scenarios. In our experiments, we segment the skull in the CBCT images into seven structures, including the maxilla,

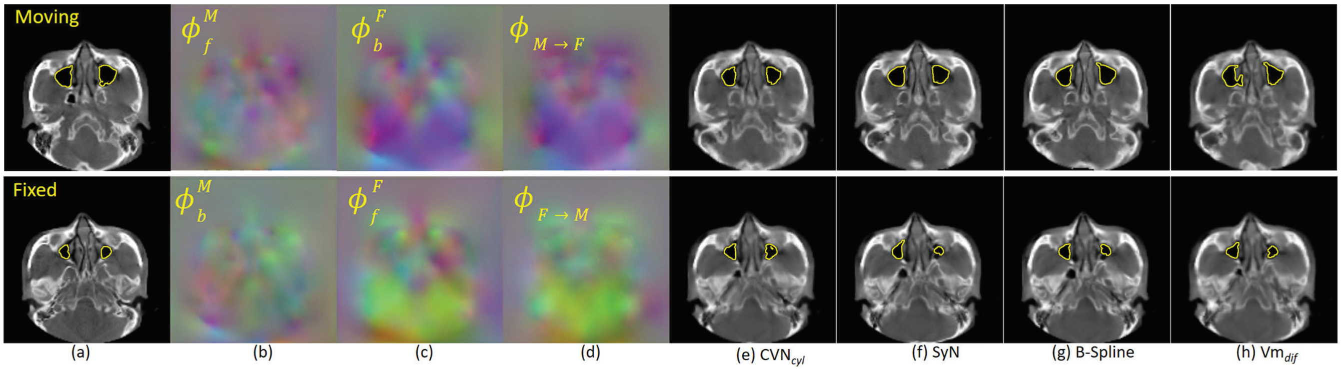

Figure 3: (a) Input volume pair (V_M, V_F). (b) and (c) are the forward and backward DVFs of V_M and V_F respectively. (d) The forward $\phi_{M\to F}$ and the backward $\phi_{F\to M}$. (e-h) are the warped volumes V_F' and V_M' using the backward $\phi_{F\to M}$ and the forward $\phi_{M\to F}$ obtained by the proposed CVN$_{cyl}$, the SyN (Avants, Epstein, and Gee 2008), the B-Spline (Rueckert et al. 1999), and theVM$_{dif}$ (Dalca et al. 2018) methods. The contours of the maxillary sinus are plotted in yellow.

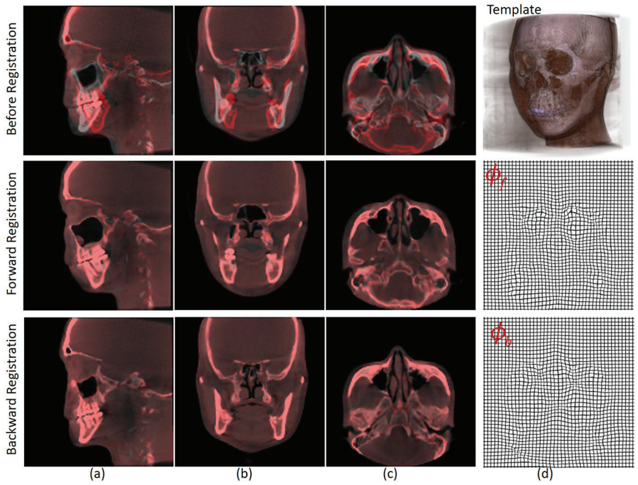

Figure 4: (a) The sagittal, (b) coronal, and (c) axial overlapping of one sampled volume pair before registration, after the forward registration using ϕ_f, and after the backward registration using ϕ_b. (d) One sampled slice of the DVFs ϕ_f and ϕ_b with the resolution of $50 \times 50 \times 50$.

the mandible, the zygoma, the frontal bone, the sphenoid bone, the occipital bone, and the temporal bone. We calculate the number of voxels with negative Jacobian determinants to assess the property of topology-preserving of DVFs, which is crucial in medical image registration.

Baseline. We compare our method with the affine registration, the B-spline-based free-form registration (Rueckert et al. 1999), the diffeomorphic SyN (Avants, Epstein, and Gee 2008) of the ANTs software package (Avants et al. 2011), and the deep learning-based Voxelmorph (VM) (Balakrishnan et al. 2018) and its diffeomorphic variant (VM$_{dif}$) (Dalca et al. 2018). We also compare the proposed CVN without the inverse constraint, the CVN$_{inv}$ with invertible bidirectional mapping (the 1st term in Eq. 4), and the CVN$_{cyl}$ with additional cycle-consistent constraints (the 2nd term in Eq. 4). The structure-aware constraints (Eq. 8) are not used in the comparison experiments except in Section 4.3.

4.1 Accuracy

Fig. 3 illustrates the bidirectional DVFs of an arbitrary moving and fixed volume pair (V_M, V_F). Each image has a pair of forward and backward DVFs regarding the template. $\phi_{M\to F} = \phi_f^M \circ \phi_b^F$, and $\phi_{F\to M} = \phi_f^F \circ \phi_b^M$. We illustrate the warped image V_M' and V_F' using the forward $\phi_{M\to F}$ and backward $\phi_{F\to M}$ obtained by the proposed CVN$_{cyl}$, the SyN (Avants, Epstein, and Gee 2008), the B-Spline (Rueckert et al. 1999), and the VM$_{dif}$ (Dalca et al. 2018) methods. The warped volume obtained by the proposed method achieves consistency with the ground truth compared with the diffeomorphic methods with and without deep learning (Dalca et al. 2018; Avants, Epstein, and Gee 2008). We visualize the contours of maxillary sinus as shown 3. The contours on the warped volumes obtained by our method are consistent with the target.

The coarse DVFs with the resolution of $50 \times 50 \times 50$ are shown in Fig. 4. The axial, coronal, and sagittal overlapping of the deformed volumes using the predicted DVFs and the targets are illustrated. We measure the MSD of the predicted DVFs from the ground truth, as shown in Table 1. The mean MSD of both the forward and the backward DVFs are below $0.25mm$.

Table 1: The MSD (mm) of the forward and the backward DVFs on the toy dataset.

	ϕ_f	ϕ_b
MSD	0.20± 0.03	0.24± 0.03

Given the DVFs, we transfer the segmentation map from one labeled volume to novel ones. Fig. 5 shows two cases of segmentation map transfer. We visualize the segmentation maps of six anatomies and plot the contours of the maxillary sinus and the mandible. The proposed method (Fig. 5(d)) is

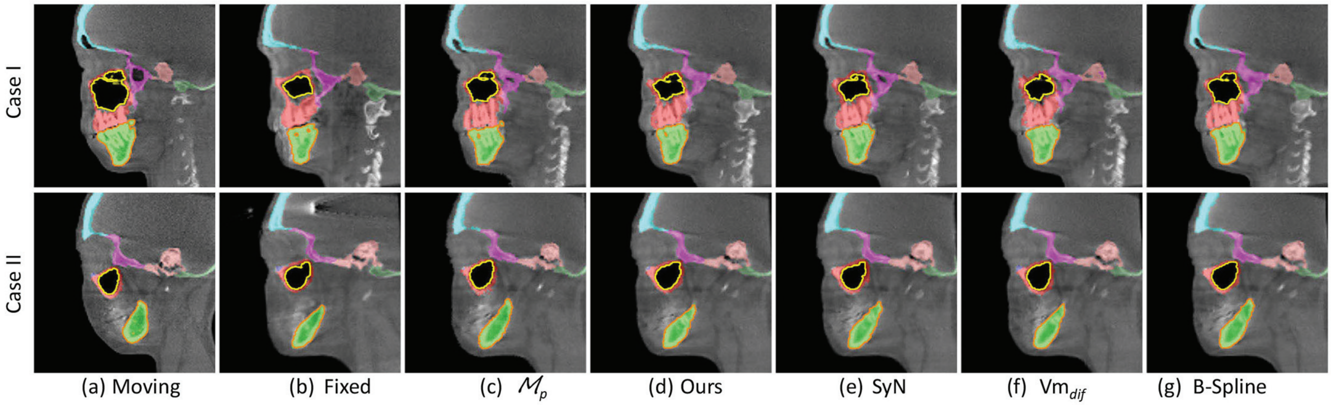

Figure 5: Comparison of segmentation map transfer. (a) Moving volume. (b) Fixed volume with ground-truth segmentation. (c-g) The segmentation map transfer using the pre-trained model $\mathcal{M}_p$, the proposed CVN$_{cyl}$, the SyN (Avants, Epstein, and Gee 2008), the VM$_{dif}$ (Dalca et al. 2018), and the B-spline-based (Rueckert et al. 1999) methods. The contours of the maxillary sinus are plotted in yellow, and contours of the mandible in orange. (Red-maxilla, green-mandible, cyan-frontal bone, magenta-sphenoid bone, dark green-occipital bone, and brown-left temporal bone.)

feasible to transfer the voxel-wise label maps. As described in Section 3.5, the proposed network is initialized by the pre-trained network using labeled synthetic dataset. The pre-trained network $\mathcal{M}_p$ also performs the volume feature extraction and DVF inference. However, the pre-trained model is not enough to predict reliable voxel-wise correspondence for the label transfer (see Fig. 5(c)). We think the reason is that the synthetic dataset is limited to cover the structural variation of the real volumes. The proposed method learned using the combinational inverse constraints is comparable with the diffeomorphic techniques (Avants, Epstein, and Gee 2008; Dalca et al. 2018) and the optimization-based free-from deformation (Rueckert et al. 1999) in the label transfer.

We report the DSC of the label transfer on seven anatomies as shown in Table 2. The proposed method achieves comparable performances with state-of-the-art diffeomorphic methods, i.e., Syn (Avants, Epstein, and Gee 2008) and diffeomorphic VM$_{dif}$ (Dalca et al. 2018). In all seven structures, the proposed CVN method gains the best performance. Note that when given the inverse constraints, the proposed CVN$_{inv}$ and CVN$_{cyl}$ significantly reduce the number of the voxels with negative Jacobian determinants from an average of 3688 to 40.6 and 7.44 respectively with small DSC costs. The proposed method is extremely faster than the traditional iterative optimization-based diffeomorphic method. Moreover, there is no need to integrate the intermediate velocity field for the final DVF as in other deep learning-based diffeomorphic registration (Dalca et al. 2018; Krebs et al. 2019).

4.2 Invertibility

Considering the inverse constraints, the concatenation of the resulted forward and backward DVFs are expected to be an identical transformation. Fig. 6 illustrates the concatenated DVFs of $\phi_f \circ \phi_b$ obtained by the CVN, the CVN$_{inv}$, and the CVN$_{cyl}$. The smaller displacements, the better. We compare with the non-diffeomorphic B-spline-based method (Rueckert et al. 1999), as well as the diffeomorphic ones including the SyN (Avants, Epstein, and Gee 2008) and VM$_{dif}$ (Dalca et al. 2018). The proposed method does not rely on the intermediate velocity vector field inference. The introduction of the inverse constraints is feasible to get an identical transformation by the concatenation of forward and backward DVFs.

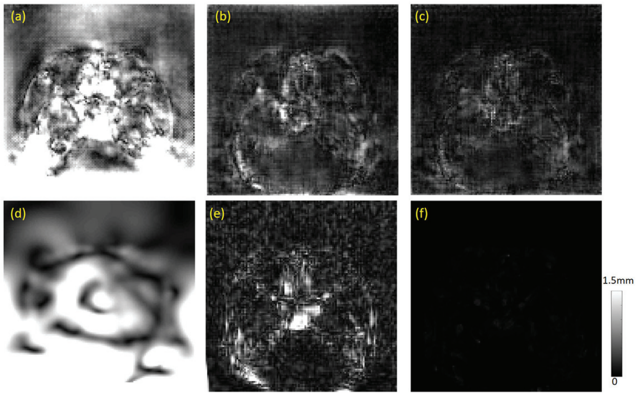

Figure 6: The concatenated displacement fields of $\phi_f \circ \phi_b$ obtained by (a) the CVN, (b) the CVN$_{inv}$, (c) the CVN$_{cyl}$, (d) the B-spline-based (Rueckert et al. 1999), (e) the VM$_{dif}$ (Dalca et al. 2018), and (f) the SyN (Avants, Epstein, and Gee 2008) methods. The smaller values, the better.

We also evaluate the consistent registration in a cycle of 3 images, as shown in Fig. 7. For a volume triplet (V_A, V_B, V_C), we compare the concatenation of DVF $\phi_{AB} \circ \phi_{BC}$ and the directional DVF ϕ_{AC}. We compare the proposed network CVN learned without the inverse constraints, the CVN$_{inv}$ with only the inverse constraints regarding the mapping to the template (the 1st term in Eq. 4), and the CVN$_{cyl}$ with the additional cycle consistency constraints. As stated in Proposition 1, the strictly invertible maps re-

Table 2: The DSC of the label transfer of seven anatomies using the proposed CVN, the CVN$_{inv}$, the CVN$_{cyl}$, as well as the affine, the B-spline-based (Rueckert et al. 1999), the SyN (Avants, Epstein, and Gee 2008), the VM (Balakrishnan et al. 2018), and the VM$_{dif}$ (Dalca et al. 2018) methods. The bottom row is the number of voxels with negative Jacobian determinants.

	Affine	B-Spline	SyN	VM	VM$_{dif}$	CVN	CVN$_{inv}$	CVN$_{cyl}$
Maxilla	0.57± 0.04	0.72 ±0.02	0.75± 0.02	0.75 ±0.01	0.74 ±0.01	**0.78**± 0.02	0.76± 0.02	0.76 ±0.02
Mandible	0.66 ±0.12	0.87 ±0.01	0.88± 0.01	0.89± 0.01	0.88 ±0.01	**0.90**± 0.01	0.88 ±0.01	0.88 ±0.01
Zygoma	0.60 ±0.09	0.79 ±0.02	0.81± 0.02	**0.83** ±0.02	0.80 ±0.02	**0.83**± 0.02	0.81± 0.02	0.81± 0.02
Frontal	0.66 ±0.05	0.80 ±0.02	0.80 ±0.02	**0.85**± 0.01	0.82± 0.01	**0.85**± 0.01	0.82 ±0.01	0.82 ±0.02
Sphenoid	0.47± 0.09	0.68 ±0.02	0.71 ±0.01	0.69 ±0.03	0.68 ±0.02	**0.73**± 0.02	0.71 ±0.02	0.70 ±0.02
Occipital	0.44 ±0.14	0.73 ±0.06	0.71± 0.10	0.71± 0.11	0.71 ±0.08	**0.80** ±0.02	0.76 ±0.04	0.76± 0.04
Temporal	0.53 ±0.08	0.75 ±0.02	0.74 ±0.04	0.79 ±0.03	0.77 ±0.02	**0.81**± 0.01	0.78± 0.02	0.78 ±0.02
$\lvert J(\phi)\rvert \leq 0$	0	2016	0	8215	500	3688	40.6	7.44

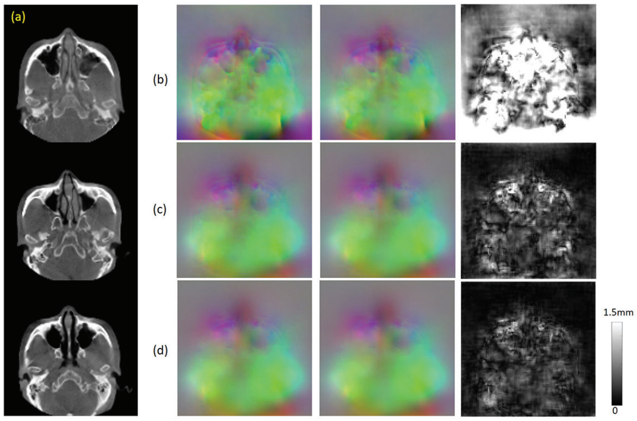

Figure 7: (a) Input volume images V_A, V_B, and V_C. The displacement fields obtained using (b) the CVN, (c) the CVN$_{inv}$, and (d) the CVN$_{cyl}$. From left to right: $\phi_{AB} \circ \phi_{BC}$, ϕ_{AC}, and the image difference of $\phi_{AB} \circ \phi_{BC} - \phi_{AC}$.

garding a shared template is enough for the cycle consistency. However, it is difficult to find strictly invertible DVFs using the parameterized functions learned by the gradient descent-based optimization. We observe that with the additional cycle consistent constraints, the concatenation of the DVFs along the path is more likely to be an identical one.

4.3 Optional Structure-aware Constraint

In our system, we introduce the optional structure-aware constraints for accurate registration of SOIs. Fig. 8 illustrates the registration of the ACB with and without the structural constraints. The contours of the sphenoidal sinus are plotted in green. We observe that the fine-grained structures in the warped image are more consistent with the target when given the additional structural constraints on the registration. Quantitative results using the mask of the sphenoid are shown in Table 3 with improved accuracies on the masked SOI. In experiments, we set the coefficient α_{str} of L_{str} to 50.

5 Conclusion

We have presented an unsupervised fully convolutional network-based framework for the cycle-consistent voxel-wise correspondence. The system takes advantages of combinational inverse constraints to enforce the invertibility of forward and backward transformation pair regarding the shared template and the cycle-consistency for the deformable registration in a volume corpus. Quantitative and qualitative results show the benefit to topology-preserving in deformable registration from involving the combinational inverse constraints. The optional bidirectional structural constraints are introduced for the registration of the fine-grained anatomies. The proposed system achieves efficient nonrigid volumetric registration and dense correspondence considering the fast evaluations of fully convolutional network-based functions, which avoids time-consuming online iterative optimization and inference of velocity volumes in other inverse-consistent registration techniques.

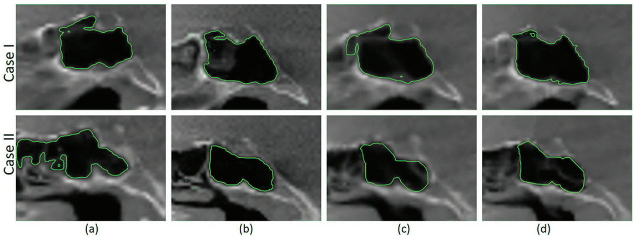

Figure 8: The registration of the ACB with and without the structural constraints. (a) Moving volume V_M. (b) Fixed volume V_F. Warped volume V'_M (c) without structural constraints, and (d) with structural constraints. The contours of the sphenoidal sinus are plotted in green.

Table 3: The DSC of the label transfer of the sphenoid bone with and without structural constraints (SC) (Eq. 8).

	w/o SC	with SC
CVN	0.73±0.02	0.74±0.02
CVN$_{inv}$	0.71±0.02	0.72±0.02

Acknowledgement. This work was supported by NKTRDP China 2017YFB1002601, NSFC 61876008, 81371192.

References

Ashburner, J., and Friston, K. J. 2000. Voxel-based morphometry the methods. *Neuroimage* 11(6):805–821.

Avants, B. B.; Tustison, N. J.; Song, G.; Cook, P. A.; Klein, A.; and Gee, J. C. 2011. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage* 54(3):2033–2044.

Avants, B.; Epstein, Clgrossman, M.; and Gee, J. 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis* 12(1):26–41.

Bajcsy, R., and Kovacic, S. 1989. Multiresolution elastic matching. *Computer Vision Graphics and Image Processing* 46(1):1–21.

Balakrishnan, G.; Zhao, A.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2018. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9252–9260.

Beg, M. F.; Miller, M. I.; Trouve, A.; and Younes, L. 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision* 61(2):139–157.

Christensen, G. E., and Johnson, H. J. 2001. Consistent image registration. *IEEE Transactions on Medical Imaging* 20(7):568–82.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, 424–432. Springer.

Dalca, A. V.; Bobu, A.; Rost, N. S.; and Golland, P. 2016. Patch-based discrete registration of clinical brain images. In *International Workshop on Patch-based Techniques in Medical Imaging*, 60–67. Springer.

Dalca, A. V.; Balakrishnan, G.; Guttag, J.; and Sabuncu, M. R. 2018. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 729–738. Springer.

de Vos, B. D.; Berendsen, F. F.; Viergever, M. A.; Staring, M.; and Išgum, I. 2017. End-to-end unsupervised deformable image registration with a convolutional neural network. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer. 204–212.

He, J., and Christensen, G. E. 2003. Large deformation inverse consistent elastic image registration. In *In Biennial International Conference on Information Processing in Medical Imaging*, 438–449. Springer.

Huang, Q.-X.; Zhang, G.-X.; Gao, L.; Hu, S.-M.; Butscher, A.; and Guibas, L. 2012. An optimization approach for extracting and encoding consistent maps in a shape collection. *ACM Transactions on Graphics (TOG)* 31(6):167.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in neural information processing systems*, 2017–2025.

Kanavati, F.; Tong, T.; Misawa, K.; Fujiwara, M.; Mori, K.; Rueckert, D.; and Glocker, B. 2017. Supervoxel classification forests for estimating pairwise image correspondences. *Pattern Recognition* 63:561–569.

Krebs, J.; Mansi, T.; Delingette, H.; Zhang, L.; Ghesu, F. C.; Miao, S.; Maier, A. K.; Ayache, N.; Liao, R.; and Kamen, A. 2017. Robust non-rigid registration through agent-based action learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 344–352. Springer.

Krebs, J.; e Delingette, H.; Mailhé, B.; Ayache, N.; and Mansi, T. 2019. Learning a probabilistic model for diffeomorphic registration. *IEEE transactions on medical imaging*.

Leow, A.; Huang, S. C.; Geng, A.; Becker, J.; Davis, S.; Toga, A.; and Thompson, P. 2005. Inverse consistent mapping in 3d deformable image registration: Its construction and statistical properties. In *Biennial International Conference on Information Processing in Medical Imaging*.

Lombaert, H.; Arcaro, M.; and Ayache, N. 2015. Brain transfer: spectral analysis of cortical surfaces and functional maps. In *International Conference on Information Processing in Medical Imaging*, 474–487. Springer.

Maes, F., et al. 1997. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging* 16(2):187–198.

Minjeong, K.; Guorong, W.; Pew-Thian, Y.; and Dinggang, S. 2012. A general fast registration framework by learning deformation-appearance correlation. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society* 21(4):1823–33.

Pluim, J. P.; Maintz, J. A.; and Viergever, M. A. 2003. Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging* 22(8):986–1004.

Rohé, M.-M.; Datar, M.; Heimann, T.; Sermesant, M.; and Pennec, X. 2017. Svf-net: learning deformable image registration using shape matching. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 266–274. Springer.

Roshni, B.; Fessler, J. A.; and Boklye, K. 2009. Accelerated nonrigid intensity-based image registration using importance sampling. *IEEE Transactions on Medical Imaging* 28(8):1208–1216.

Rueckert, D.; Sonoda, L. I.; Hayes, C.; Hill, D. L.; Leach, M. O.; and Hawkes, D. J. 1999. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging* 18(8):712–721.

Sokooti, H.; de Vos, B.; Berendsen, F.; Lelieveldt, B. P.; Išgum, I.; and Staring, M. 2017. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 232–239. Springer.

Sotiras, A.; Davatzikos, C.; and Paragios, N. 2013. Deformable medical image registration: A survey. *IEEE transactions on medical imaging* 32(7):1153.

Thirion, J. P. 2011. Image matching as a diffusion process: an analogy with maxwell' demons. *Medical Image Analysis* 2(3):243.

Vercauteren, T.; Pennec, X.; Perchant, A.; and Ayache, N. 2009. Diffeomorphic demons: Efficient non-parametric image registration. *Neuroimage* 45(1):61–72.

Wei, L.; Cao, X.; Wang, Z.; Gao, Y.; Hu, S.; Wang, L.; Wu, G.; and Shen, D. 2017. Learning-based deformable registration for infant mri by integrating random forest with auto-context model. *Medical Physics* 44(12).

Yang, X.; Kwitt, R.; Styner, M.; and Niethammer, M. 2017. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage* 158:378–396.

Zhou, T.; Krahenbuhl, P.; Aubry, M.; Huang, Q.; and Efros, A. A. 2016. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 117–126.