Medical Image Analysis 99 (2025) 103351

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

MUsculo-Skeleton-Aware (MUSA) deep learning for anatomically guided head-and-neck CT deformable registration

Hengjie Liu [a,b], Elizabeth McKenzie [c], Di Xu [d,e], Qifan Xu [d,e], Robert K. Chin [b], Dan Ruan [a,b], Ke Sheng [d,e,*]

[a] Physics and Biology in Medicine Graduate Program, University of California Los Angeles, Los Angeles, CA, USA
[b] Department of Radiation Oncology, University of California Los Angeles, Los Angeles, CA, USA
[c] Department of Radiation Oncology, Cedars-Sinai Medical Center, Los Angeles, CA, USA
[d] UCSF/UC Berkeley Graduate Program in Bioengineering, University of California San Francisco, San Francisco, CA, USA
[e] Department of Radiation Oncology, University of California San Francisco, San Francisco, CA, USA

ARTICLE INFO

Keywords:
Deformable image registration
Deep learning
Anatomical constraint
Head and neck CT

ABSTRACT

Deep-learning-based deformable image registration (DL-DIR) has demonstrated improved accuracy compared to time-consuming non-DL methods across various anatomical sites. However, DL-DIR is still challenging in heterogeneous tissue regions with large deformation. In fact, several state-of-the-art DL-DIR methods fail to capture the large, anatomically plausible deformation when tested on head-and-neck computed tomography (CT) images. These results allude to the possibility that such complex head-and-neck deformation may be beyond the capacity of a single network structure or a homogeneous smoothness regularization. To address the challenge of combined multi-scale musculoskeletal motion and soft tissue deformation in the head-and-neck region, we propose a MUsculo-Skeleton-Aware (MUSA) framework to anatomically guide DL-DIR by leveraging the explicit multi-resolution strategy and the inhomogeneous deformation constraints between the bony structures and soft tissue. The proposed method decomposes the complex deformation into a bulk posture change and residual fine deformation. It can accommodate both inter- and intra- subject registration. Our results show that the MUSA framework can consistently improve registration accuracy and, more importantly, the plausibility of deformation for various network architectures. The code will be publicly available at https://github.com/HengjieLiu/DIR-MUSA.

1. Introduction

Deformable image registration (DIR) plays a crucial role in medical image analysis, with applications in diagnosis, image-guided surgery, and radiation therapy. DIR aims to find the anatomically accurate spatial correspondence between a pair of fixed and moving images, represented as a deformation field. DIR can be performed for the same subject (intra-subject) between different time points and/or imaging modalities, or among different subjects (inter-subject). It is typically formulated as a regularized optimization problem that is iteratively solved for a given image pair. The different combinations of cost functions, regularization methods, and optimization algorithms have led to different iterative DIR methods (Sotiras et al., 2013). These DIR methods have enabled and improved imaging information synthesis in medical research and practice. As iterative architectures mature, their performance has plateaued with varying usability depending on the required accuracy and problem complexity (Li et al., 2017). Iterative DIR methods often require manual parameter tuning on a specific dataset to achieve optimal performance, considerably limiting their robustness and generalizability. Moreover, iterative methods are typically computationally demanding and time-consuming, restricting applications that require real-time outputs.

Recent advances in deep learning (DL) have offered a new perspective for solving DIR problems. Unlike conventional iterative methods that derive the voxel-level correspondence solely based on a single pair of images, deep-learning-based DIR (DL-DIR) methods learn the statistical correspondence between images within a large training set, and then apply this knowledge to unseen imaging pairs. Its real-time inference capability is essential for time-sensitive tasks such as motion tracking and prediction, facilitating image guidance in surgery and in radiation therapy. However, the absence of ground-truth deformation

* Corresponding author at: Department of Radiation Oncology, University of California San Francisco, San Francisco, CA, USA.
E-mail address: ke.sheng@ucsf.edu (K. Sheng).

https://doi.org/10.1016/j.media.2024.103351
Received 29 September 2023; Received in revised form 5 June 2024; Accepted 16 September 2024
Available online 21 September 2024

poses challenges to supervised DL-DIR. The performance of DIR has been indirectly assessed with intensity-matching metrics and contour agreement. Landmark matching can provide a more accurate evaluation when available, yet it is often limited by sparse and site-specific labeling. Consequently, unsupervised and weakly-supervised learning have emerged as the most popular and effective methods for this task. Following the success of VoxelMorph (Balakrishnan et al., 2019), different convolutional neural network (CNN) designs, mostly U-Net (Ronneberger et al., 2015) variants, have been proposed to tackle the registration challenge at various anatomical sites. More recently, Transformers have been introduced to medical image registration (Chen et al., 2021, 2022) and demonstrated improved performance compared to CNN baselines.

Nevertheless, obtaining accurate and anatomically plausible deformations for head-and-neck computed tomography (CT) registration remains challenging. On the one hand, the large magnitude of motion and high degrees of motion freedom in the head-and-neck region can pose challenges for DIR methods searching for a local match. On the other hand, the deformation is a complex superposition of musculoskeletal motion and soft tissue deformation, which is beyond the descriptive capacity of a homogeneous smoothness regularization typically employed in DIR. In conventional registration approaches, hierarchical or multiresolution strategies are commonly used to avoid local minima and to simultaneously recover the global large deformation and the local detailed deformation (Klein et al., 2010; Lester and Arridge, 1999). Similar ideas have been adopted by quite a few deep learning methods, where image pyramids are built and the deformation is optimized in a coarse-to-fine manner (Eppenhof et al., 2020; Hering et al., 2021; Kang et al., 2022; Mok and Chung, 2020). However, many other methods still rely solely on the hierarchical architecture intrinsic to the neural networks (Hering et al., 2021), such as U-Net (Ronneberger et al., 2015) and Swin Transformer (Liu et al., 2021). We refer to these two approaches as explicit and implicit multiresolution methods, respectively. To further address the complex and heterogeneous deformation in the head-and-neck region, previous studies have attempted to integrate anatomical or biomechanical properties into conventional registration methods. Kim et al. (2013) used a rigidity constraint on bony structures for intra-subject cone-beam CT (CBCT) to CT registration. More complex methods employed biomechanical models, including finite element methods (FEM) (Al-Mayah et al., 2010; Kim et al., 2016) and kinematic motion models (du Bois d'Aische et al., 2005a; Neylon et al., 2014; Teske et al., 2017). Biomechanical model-based registration offers the potential for more accurate and physically plausible deformations, yet it is limited by the need for precise tissue property modeling and high computational demands. Furthermore, it is only suitable for intra-subject registration. To this date, only a few deep-learning-based studies have been proposed for head-and-neck CT registration (Lei et al., 2022; Li et al., 2023b; Liang et al., 2021), while none of them are purposefully designed to address the unique characteristics of head-and-neck deformation. The head-and-neck region is also absent from the recent Learn2Reg challenge (Hering et al., 2023), which is by far the most comprehensive evaluation challenge in medical image registration.

In this work, we propose a MUsculo-Skeleton-Aware (MUSA) framework for head-and-neck CT registration. The complex deformation in the head-and-neck region can be decomposed into musculoskeletal motion and residual tissue deformation. However, a perfect decomposition demands comprehensive and individualized biomechanical modeling that is impractical and likely unnecessary. Alternatively, we propose a relaxed decomposition of the deformation into a bulk posture change and residual fine deformation, tackled by two registration networks sequentially. First, a posture correction network (Pos-Net) is employed to align the posture between the fixed and moving images. The differences in posture are predominantly caused by musculoskeletal motion, which includes head pitching, neck flexion and extension, and jaw movement. To achieve that, we propose a MUSA loss function, which encourages the local affine motion within each bony structure via increased bending energy regularization. We design it to be affine instead of strictly rigid to work for both inter-subject and intra-subject registration. In the second stage, a refinement network (Ref-Net) employing a standard homogeneous smoothness constraint is used to account for any residual fine-scale deformation. Given the distinct characteristics of the decomposed deformations, the Pos-Net takes low-resolution inputs, while the Ref-Net operates on full-resolution images, forming a two-level image pyramid as the multiresolution scheme. By leveraging this explicit multiresolution strategy and the inhomogeneous deformation constraints between the bony structures and soft tissue, our method achieves more anatomically realistic deformation estimation in the head-and-neck region with improved registration accuracy and is more interpretable. The proposed framework is compatible with most network architectures used for image registration.

1.1. Contribution

The main contributions of this study are summarized as follows:

- We propose a two-stage head-and-neck CT registration framework that decomposes the complex deformation into a bulk posture change and residual fine deformation. We propose a MUSA loss using spatially variant regularization on soft tissue and bony structures to anatomically guide the registration during the posture correction stage, which addresses the musculoskeletal motion and provides a better initial alignment to ease the burden of subsequent registration.
- The proposed framework is compatible with various network architectures. We perform comprehensive experiments to demonstrate the proposed framework can consistently achieve better registration accuracy and, more importantly, improve the anatomical plausibility of deformations. We demonstrate its performance on both inter- and intra-subject registration tasks.
- We demonstrate the importance of both explicit multiresolution strategy and anatomical guidance to ensure anatomically plausible deformations.
- We emphasize the need for a more comprehensive evaluation of deformable image registration, particularly focusing on deformation plausibility.

2. Related work

In this section, we briefly review (i) the problem formulation of deformable image registration, (ii) methods that incorporate mathematical, physical, and anatomical priors for image registration, (iii) methods for handling large motion in image registration, and (iv) head-and-neck CT registration.

For more comprehensive reviews of medical image registration, the readers can refer to Maintz and Viergever (1998); Sotiras et al. (2013) for conventional methods, and Boveiri et al. (2020); Chen et al. (2023b); Fu et al. (2020); Haskins et al. (2020) for deep-learning-based methods, respectively.

2.1. Deformable Image registration

The goal of deformable image registration (DIR) is to find a transformation ϕ that establishes the spatial correspondence between a fixed image ($I_f(\boldsymbol{x})$ or f) and a moving image ($I_m(\boldsymbol{x})$ or m). $I_f(\boldsymbol{x})$ and $I_m(\boldsymbol{x})$ are n-dimensional images defined on their own spatial domain: $\Omega_f,\ \Omega_m \subset \mathrm{R}^n$. The transformation ϕ can be parameterized by various methods but can be typically represented as a dense displacement vector field (DVF) $\boldsymbol{u}(\boldsymbol{x})$:

$$\phi = Id + \boldsymbol{u}(\boldsymbol{x}), \tag{1}$$

where Id is the identity coordinate transform and $\boldsymbol{u}(\boldsymbol{x})$ specifies the vector offset from coordinates of $I_f(\boldsymbol{x})$ to coordinates of $I_m(\boldsymbol{x})$. The

deformed image ($I_m(\boldsymbol{x})\circ\phi$) that aligns with the fixed image is produced by applying ϕ to the moving image. In the rest of the paper, we use f and m to denote $I_f(\boldsymbol{x})$ and $I_m(\boldsymbol{x})$ for simplicity. The images are represented as discrete matrices in our study.

In supervised deep learning methods, neural networks are trained to predict the deformation (ϕ or $\boldsymbol{u}$) directly. This requires the ground-truth deformation, which is hard to estimate accurately in real-world applications. Common workarounds involve training on synthetic deformations (Eppenhof et al., 2018; Sokooti et al., 2017) or using deformations generated by conventional registration methods (Cao et al., 2018; Sentker et al., 2018; Yang et al., 2017). In both cases, the registration performance is limited by the ground-truth provided, which either lacks realistic and diverse anatomical information in synthetic cases or is curtailed by the capabilities of conventional methods. As a result, the research focus has shifted to unsupervised (or self-supervised) approaches.

Alternatively, DIR can be formulated as a regularized optimization problem, which is employed by both unsupervised deep learning methods and conventional iterative methods:

$$\widehat{\phi} = \underset{\phi}{\operatorname{argmin}}\ L_{sim}(f,\ m\circ\phi) + \lambda R(\phi). \tag{2}$$

L_{sim} is the similarity metric used to quantify the spatial alignment between the fixed image (f) and the deformed image ($m\circ\phi$). The commonly used similarity metrics include mean squared error (MSE), normalized cross-correlation (NCC), and mutual information (MI). However, DIR is a highly underdetermined problem, requiring regularization ($R(\phi)$) to stabilize solutions and avoid undesired local minima. The regularization injects prior knowledge of the deformation, such as smoothness, to approach a more feasible solution. λ is the hyperparameter that controls the trade-off between image matching and deformation regularity.

In conventional iterative methods, Eq. (2) is solved for a single image pair, whereas in unsupervised deep learning approaches, it is optimized over a large training set in the sense of expectation. Unsupervised training via backpropagation was enabled by differentiable image warping, which was first implemented in the spatial transformer network (STN) (Jaderberg et al., 2015). de Vos et al. (2017) first used this formulation for CNN-based 2D registration, and the VoxelMorph paper (Balakrishnan et al., 2019) popularized this idea with a demonstration of 3D brain registration using a U-Net architecture (Ronneberger et al., 2015). The majority of subsequent research used different variations of U-Net (Hering et al., 2021; Kang et al., 2022; Kim et al., 2021; Mok and Chung, 2020). Recently, Transformers (Liu et al., 2021; Vaswani et al., 2017) have gained popularity in DIR (Chen et al., 2021, 2022; Shi et al., 2022) after their success on many computer vision tasks.

Auxiliary anatomical information, such as segmentation overlap and landmark correspondence, can be incorporated into the loss function to improve anatomical matching. These methods are categorized as weakly-supervised or semi-supervised approaches. Enforcing segmentation overlap using the Dice loss has been proposed to guide the registration network anatomically (Balakrishnan et al., 2019; Hu et al., 2018b). It has become a common component in various registration tasks (Hering et al., 2023). However, the segmentation overlap can still be ambiguous within the organ boundary. Additionally, the correlation between segmentation overlap and registration accuracy can vary. Only small and localized regions show high correlations as they approximate point landmarks more closely (Rohlfing, 2012). On the other hand, landmark correspondence provides a more accurate evaluation at specific anatomical points. It was exploited to improve intra-subject lung registration (Heinrich and Hansen, 2022; Hering et al., 2021). However, its application is mostly limited to intra-subject lung registration, as defining landmarks densely and unambiguously in other anatomical regions or inter-subject settings poses significant challenges.

2.2. Mathematical, physical, and anatomical priors for deformation regularization

DIR driven by image similarity and tissue overlap does not guarantee anatomically feasible deformations, as the algorithms are susceptible to anatomical ambiguity, image noise, and insufficient tissue contrast (Rohlfing, 2012). Incorporating prior information, including mathematical or physical properties of the deformation as well as anatomical knowledge, has proven critical to address this issue.

Desired mathematical or physical properties of the underlying deformation field are commonly enforced by designing a regularization energy function. Smoothness regularization has been universally used to ensure a physically realistic deformation. This is typically achieved by penalizing the first-order or second-order gradient of the deformation. Another common regularizer is the Jacobian determinant of the deformation, which can be used to preserve topology (i.e., prevent folding) (Christensen and Johnson, 2001; Rueckert et al., 2006), and control the volume change (Dauguet et al., 2009; Ruhaak et al., 2017). Diffeomorphism is also appealing as it guarantees a continuous and invertible deformation field. Dalca et al. (2019) incorporated the scaling-and-squaring strategy (Arsigny et al., 2006) to generate near diffeomorphic deformation from stationary velocity fields. Furthermore, different approaches have been explored to encourage inverse consistency (Greer et al., 2021; Kim et al., 2021; Zhang, 2018).

For smoothness regularization, most studies adopt a uniform weight across the entire deformation field, which is typically determined by hyperparameter tuning. However, this assumption can be suboptimal as different anatomical regions can present different levels of deformation regularities determined by the underlying anatomical or biomechanical properties (Chen et al., 2023b). A variety of studies have developed regularizers that can vary in space with conventional optimization-based schemes (e.g., Gerig et al., 2014; Kabus et al., 2006; Pace et al., 2013; Vialard and Risser, 2014). More recently, several deep-learning-based studies have proposed to learn spatially variant regularization directly from data. Niethammer et al. (2019) used a learnable Gaussian smoothing kernel map, parameterized by a neural network, to represent varying smoothness levels. Recent studies on brain registration introduced conditional networks or subnetworks to achieve spatially variant regularization. These methods further support inference time adaptation, eliminating the need for hyperparameter tuning (Chen et al., 2023a; Wang et al., 2023).

Anatomical and biomechanical information can provide further guidance for obtaining a feasible deformation. For instance, the presence of stiff anatomical structures motivates the incorporation of rigidity constraints in conventional methods (Loeckx et al., 2004; Ruan et al., 2006; Staring et al., 2007). Jian et al. (2022) incorporated rigidity constraint in deep-learning-based spine registration. More sophisticated prior knowledge can be derived from biomechanical models, with finite element methods (FEMs) being popular for modeling tissues with varying material properties (Sotiras et al., 2013). Biomechanical model-based registration has been applied to various anatomical sites (e.g. Bharatha et al., 2001; Rajagopal et al., 2007; Sermesant et al., 2003; Werner et al., 2009) and also multi-organ registration (Brock et al., 2005; He et al., 2023). Recently, many deep learning methods have incorporated deformation fields generated by biomechanical models for supervision or regularization (e.g. Fu et al., 2021; Hu et al., 2018a; Qin et al., 2020; Zhang, 2021).

2.3. Handling large motion in image registration

Large deformation presents a significant challenge for DIR. In conventional iterative methods, multiresolution or hierarchical strategies have become standard to avoid local minima, speed up convergence, reduce foldings, and recover global and local motion simultaneously (Bajcsy and Kovačič, 1989; Klein et al., 2010; Lester and Arridge, 1999; Schnabel et al., 2001). This typically involves forming an image pyramid

through downsampling and/or blurring, with deformation estimated from a coarser scale and incrementally refined to higher resolutions (Klein et al., 2010).

However, not all deep-learning-based methods adopt this strategy, likely because the commonly used network architectures, such as U-Net (Ronneberger et al., 2015) and Swin Transformers (Liu et al., 2021), are considered inherently hierarchical (Hering et al., 2021), where the network encoder generates multiresolution features by spatial downsampling, and the decoder restores the original resolution. We refer to these methods as the implicit multiresolution approach. However, these methods typically predict deformations only at the finest resolution, which can trap optimization in local minima due to the ill-posed nature of DIR, where many possible transformations can result in comparable similarity matching (Mok and Chung, 2020). This issue can be further deteriorated by the limited effective receptive field (ERF) of CNNs (Chen et al., 2022).

Recently, more studies have integrated the explicit multiresolution approach from conventional methods and shown improved handling of large motions. Eppenhof et al. (2020) proposed a novel training strategy by progressively growing the U-Net from low to high resolution. Hering et al. (2021, 2019) proposed to sequentially deform the moving image from coarser to finer scales to address the large motion in lungs. LapIRN (Mok and Chung, 2020) applied Laplacian image pyramids to refine deformation progressively via addition until the finest scale, demonstrating strong performance across multiple registration tasks (Hering et al., 2023). Dual-PRNet (Hu et al., 2019; Kang et al., 2022) developed a dual-stream pyramid for feature encoding and separated the feature learning and deformation estimation process. The deformation predicted from a coarser level was used to align the features from a higher resolution, followed by deformation refinement in the higher resolution.

Alternatively, the cascaded-based approach, where each network handles a small fraction of the deformation, has been explored to handle large motion (Zhao et al., 2020). However, cascading the same network leads to high computation costs and long inference times. More recently, Hu et al. (2022) demonstrated that the pyramid-based and cascaded-based methods can be combined in a recursive decomposition framework.

Enlarging the ERF has also been proven to be helpful in handling large motion, especially with the recent introduction of Transformers. Chen et al. (2022) proposed a hybrid Swin Transformer (Liu et al., 2021) and CNN architecture called TransMorph, achieving state-of-the-art performance on several medical image registration tasks. The capacity for long-range dependency modeling and larger ERFs enabled by the self-attention mechanism were claimed to be a major advantage of using Transformers over CNNs for image registration (Chen et al., 2022; Li et al., 2023a). However, RepLKNet, proposed by Ding et al. (2022), suggested that CNNs could also achieve larger ERFs with larger convolutional kernels. Jia et al. (2022) further applied this idea for image registration using a large kernel U-Net (LK-U-Net) and demonstrated performance comparable to TransMorph on several brain datasets.

2.4. Head-and-neck CT registration

Deformable registration of head-and-neck CT is a unique problem due to the complex superposition of musculoskeletal motion and residual soft tissue deformation. The head pitching, neck flexion and extension, along with jaw movement, can result in large posture differences. Rigidity constraints on the bony structures, such as the skull, mandible, and vertebrae, can be used to guide intra-subject registration. Furthermore, biomechanical models using finite element methods and kinematic motion models have been proposed to model the head-and-neck motion.

Kim et al. (2013) proposed a novel rigidity constraint based on distance preservation as an alternative to the rigidity constraint proposed by (Staring et al., 2007), as the latter lacks the ability to separately preserve the rigidity of multiple objects in close proximity in the head-and-neck region. du Bois d'Aische et al. (2005b, 2005a) proposed registering the articulated rigid bones first, then propagating the deformation to the soft tissue using a linear elastic model. Neylon et al. (2014) developed a GPU-based biomechanical model, where the muscle and soft tissue structures were modeled as a mass-spring model, allowing them to deform along with the articulating skeletal structure to simulate posture changes. In Al-Mayah et al. (2010), patient-specific 3D FEMs were developed to align CT with cone-beam CT. The model included seven vertebrae (C1 to C7), the mandible, the larynx, the parotid glands, the tumor, and the body. Several studies used FEM to evaluate registration accuracy in radiation therapy clinics (Kim et al., 2016; McCulloch et al., 2019; Teske et al., 2017). Biomechanical model-based methods can improve registration accuracy and plausibility when tissue properties and boundary conditions are well-understood. However, accurately measuring these parameters in vivo is challenging and also patient-specific, making the method sensitivity to inaccurate assumptions (Chi et al., 2006; Hipwell et al., 2016). Additionally, these methods require substantial computational resources and are generally limited to intra-subject registration due to variability in biomechanical properties among individuals (Sotiras et al., 2013).

Only a few deep-learning-based methods have been proposed to address head-and-neck CT registration. Liang et al. (2021) employed 5 cascades of VoxelMorph for intra-subject head-and-neck CT registration for auto segmentation, achieving accuracy only comparable to that of Elastix (Klein et al., 2010). Also, cascading 5 networks can cause significant computation and memory burdens. Lei et al. (2022) introduced a dual feasible network for intra-subject head-and-neck CT registration, but the specific clinical setting with small initial misalignment of target registration error (TRE) around 4mm was inadequate to demonstrate its effectiveness in handling large motion. Li et al. (2023b) combined deep-learning-based feature extraction and convex optimization for multiple CT registration tasks including a head-and-neck dataset. The methodology of these studies remains generic and does not consider the distinct deformation patterns or anatomical properties in head-and-neck regions.

3. Methods

In this section, we describe the details of the proposed MUSA framework.

3.1. MUSA loss function

A standard registration loss function (i.e., Eq. (2)) employing MSE as similarity metric and a homogeneous smoothness regularization with bending energy (BE) can be written as:

$$Loss = MSE(f, m\circ\phi) + \lambda \sum_{\boldsymbol{r}\in\Omega_f} BE(\phi(\boldsymbol{r})). \tag{3}$$

Bending energy is defined as:

$$BE = \frac{1}{V}\int\int\int_V \sum_{i=x,y,z}\left[\left(\frac{\partial^2 u_i}{\partial x^2}\right)^2 + \left(\frac{\partial^2 u_i}{\partial y^2}\right)^2 + \left(\frac{\partial^2 u_i}{\partial z^2}\right)^2 + 2\left(\frac{\partial^2 u_i}{\partial xy}\right)^2 + 2\left(\frac{\partial^2 u_i}{\partial yz}\right)^2 + 2\left(\frac{\partial^2 u_i}{\partial xz}\right)^2\right] dxdydz, \tag{4}$$

where u_i is the component of the displacement vector $\boldsymbol{u}$.

We select MSE as the similarity metric as it suits well for mono-modal registration and the quantitative nature of CT images. It is also more computationally efficient compared with other metrics. Other similarity metrics, such as NCC, could be used as well. Bending energy is a commonly used smoothness regularizer (Rueckert et al., 1999). It penalizes the second-order derivative of the deformation field. As an additional advantage, it zeros out any linear component so that the global affine registration can be integrated within the deformable registration without increasing the loss (Chen et al., 2023b). Thus, additional affine pre-alignment is not necessary (Ding and Niethammer, 2022; Fischer and Modersitzki, 2003).

To address the complex head-and-neck deformation, we propose a MUsculo-Skeleton-Aware (MUSA) loss function, with increased bending energy regularization on bony structures compared to soft tissue:

$$Loss = MSE(f, m\circ\phi) + \lambda\left(\sum_{\boldsymbol{r}\in S} BE(\phi(\boldsymbol{r})) + \alpha\sum_{\boldsymbol{r}\in B} BE(\phi(\boldsymbol{r}))\right). \quad (5)$$

S is the soft tissue mask and B is the bone mask, defined on the deformed image ($m \circ \phi$) coordinate. λ controls the overall strength of regularization, and α controls the relaxed rigidity of bony structure relative to that of the soft tissue. Note that when $\alpha = 1$, Eq. (5) reduces to the standard loss function (i.e., Eq. (3)).

A bending energy of zero corresponds to a pure affine transformation. So, the MUSA loss with a large α value encourages local affine motion in each individual bone. On the one hand, it can be viewed as applying an inhomogeneous or spatially variant bending energy regularization on the deformation field. But instead of learning the weights directly from data (Chen et al., 2023a; Wang et al., 2023), we adopt a simpler but more faithful approach by assigning weights to the two distinct tissue types in CT images based on well-understood anatomical knowledge. The weights are determined by the hyperparameter tuning described in Section 4.3. On the other hand, the formulation in Eq. (5) is a relaxed version of the rigidity loss proposed by Staring et al. (2007), where orthonormality and properness conditions are used in addition to affine constraint to enforce rigidity. Since orthonormality and properness conditions both encourage volume preservation, they are not suitable for inter-subject registration. By using the affine constraint alone, we allow the volume of bony structures to adjust in the registration process if needed.

3.2. MUSA framework

The relaxed rigidity constraint in MUSA loss utilizes differential regularization of bony structures and soft tissue to promote deformation plausibility. However, the regularization can be too restrictive and will prevent an acceptable spatial alignment between the deformed and fixed images in the presence of large deformation. This is particularly problematic for inter-subject registration, where the deformation in soft tissue can be substantial and the bone shapes can differ significantly among individuals. Therefore, achieving a balance between acceptable

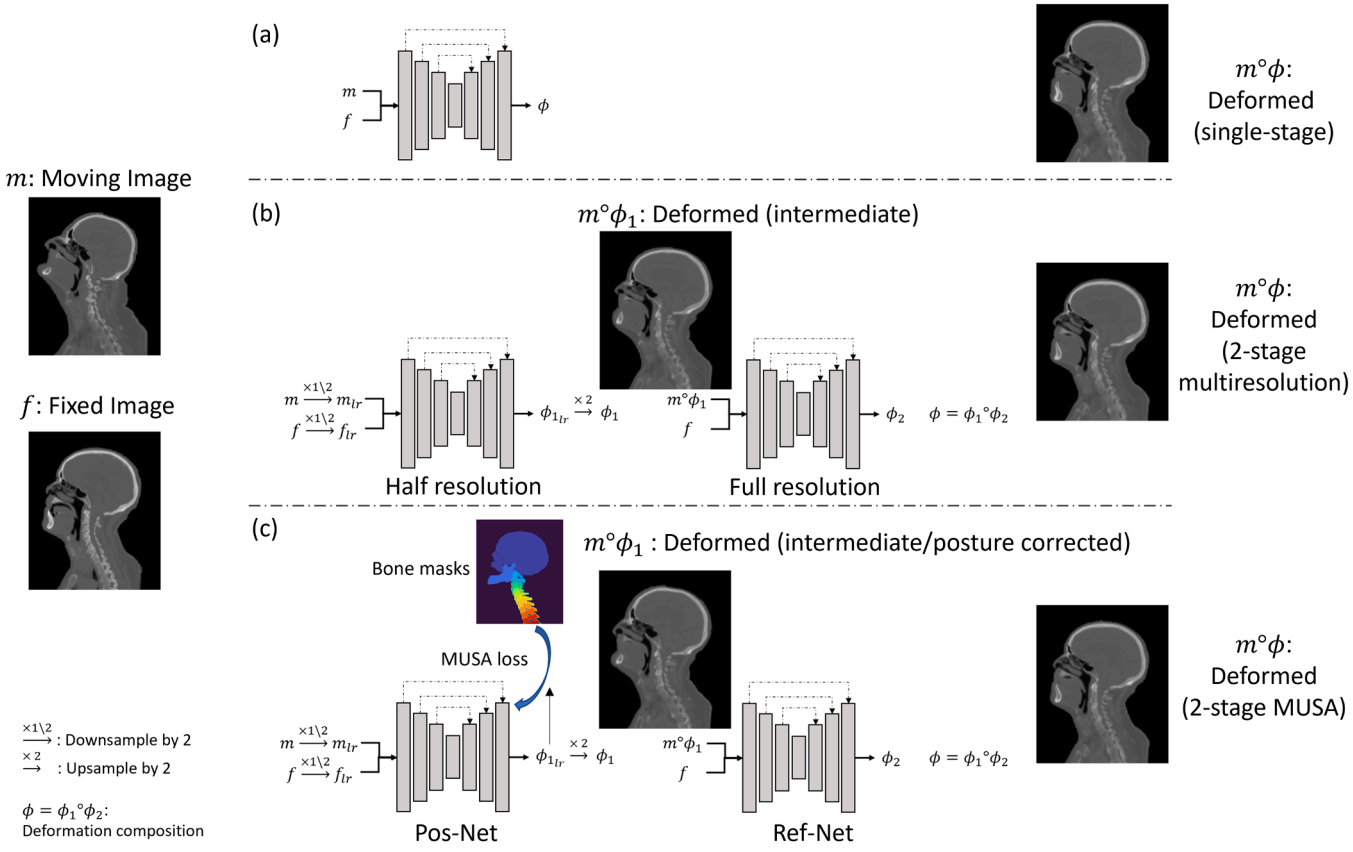

Fig. 1. The proposed MUSA registration framework (c) and baselines (a, b). The MUSA framework (c) employs a two-stage approach utilizing a two-level image pyramid for explicit multiresolution setup. In the first stage, the Pos-Net takes the low-resolution image pair (m_{lr} and f_{lr}) as inputs and is trained with the MUSA loss calculated using bony segmentations. The first-stage deformation (ϕ_1) warps the moving image (m) to a posture-corrected deformed image ($m\circ\phi_1$) as an intermediate state. In the second stage, Ref-Net takes the full-resolution pair ($m\circ\phi_1$ and f) and predicts residual deformation (ϕ_2), with the standard homogeneous smoothness regularization. The final deformation is the composition of the two stages ($\phi = \phi_1\circ\phi_2$), and the final deformed image is ($m\circ\phi$). The two-stage multiresolution baseline (b) is similar to the two-stage MUSA framework, except that both stages use homogeneous smoothness regularization. The single-stage baseline (a) directly predicts the deformation on full-resolution image pairs with homogeneous smoothness regularization. The neural networks in the figure can adopt most architectures designed for deep-learning-based DIR.

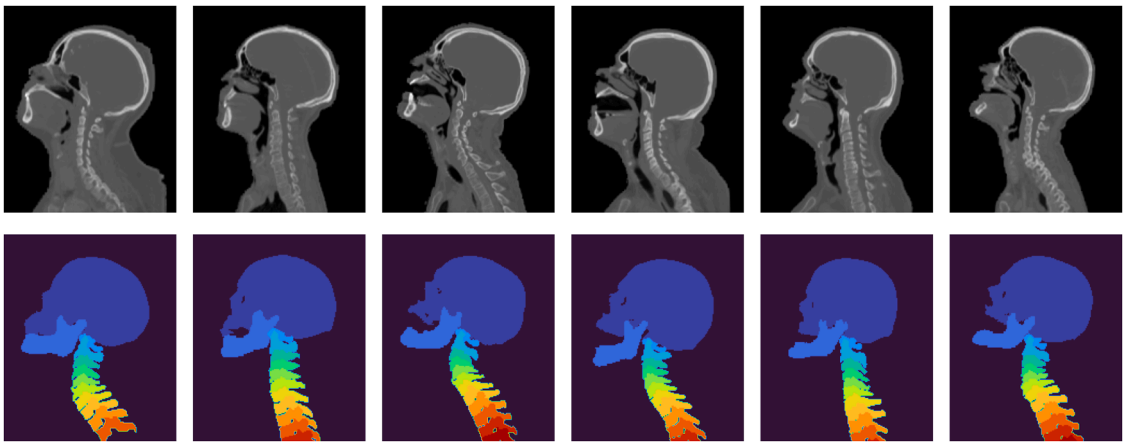

Fig. 2. Examples of head-and-neck CT images and their corresponding bone segmentations used for MUSA loss calculation. The first row presents the central slice of the CT images in coronal views. The second row displays the bone segmentations rendered using maximum intensity projection (MIP).

intensity matching and a plausible deformation representation within a single stage—utilizing a unified network and loss function—can be challenging due to the complexity of deformations in the head-and-neck region. To address this issue, we propose a two-stage decomposition approach as our MUSA registration framework.

The entire MUSA framework is illustrated in Fig. 1 (c). We aim to decompose the complex deformation in the head-and-neck region into musculoskeletal motion and soft tissue deformation, leveraging both the explicit multiresolution strategy and inhomogeneous deformation constraints between bony structures and soft tissue. Without depending on the hard-to-obtain individual subject's tissue biomechanical properties, we propose a relaxed deformation decomposition with two stages: a posture correction step using Pos-Net and a fine-scale refinement step with Ref-Net. The Pos-Net is trained with MUSA loss (i.e., Eq. (5)) and the Ref-Net is trained with the standard loss function (i.e., Eq. (3)). The two-stage decomposition is naturally integrated with a two-level multiresolution approach. This stems from the fact that the bulk posture change is a global motion, while the residual tissue deformation occurs on a finer resolution scale. During the posture correction stage, MUSA loss provides anatomical guidance by encouraging local close-to-affine motion of the bony structures. It promotes a more plausible motion pattern within each bony structure, and the effect can further extend to surrounding soft tissue due to the smoothness constraint. This guides the registration to align the posture differences of two head-and-neck images, predominantly driven by musculoskeletal motion, including head pitching, neck flexion and extension, and jaw movement. The stringent regularization posed by MUSA loss would prevent a perfect match between the deformed and fixed images, yet it provides a better-conditioned intermediate registration without unrealistically deforming the bony structures, which could otherwise lead the optimization process to local minima, resulting in implausible deformations. As a result, it significantly eases the burden of subsequent registration. In the second stage, the Ref-Net with regular homogeneous smoothness regularization is applied to account for any residual deformation, including residual soft tissue deformation and anatomical mismatch in inter-subject scenarios. The final deformation field is the composition of the two deformation outputs from Pos-Net and Ref-Net.

Fig. 1 also shows the two baseline frameworks constructed to compare with the proposed MUSA framework: (a) single-stage baseline and (b) two-stage multiresolution baseline. Both baseline frameworks use the standard loss function with homogeneous regularization (i.e., Eq. (3)). The two-stage multiresolution baseline (b) is constructed with an identical explicit multiresolution scheme as the two-stage MUSA approach to ensure a fair comparison, especially for architectures lacking explicit multiresolution modeling.

3.3. Bony structure segmentation

The proposed MUSA loss relies on the segmentation of the bony structures, which can be obtained with acceptable performance via deep-learning-based auto segmentation. A three-stage coarse-to-fine localization and segmentation network was used to acquire vertebrae segmentation (Payer et al., 2020; Sekuboyina et al., 2021). We used a shape-constrained segmentation network (Tong et al., 2019) to segment the mandible. Following the mandible and vertebrae segmentation, a skull mask was obtained via thresholding. Fig. 2 presents examples of segmented bony structures used in the study. It also demonstrates the substantial posture variance and large deformation that pose significant challenges for head-and-neck CT registration. The bony segmentations are only used in loss calculation during the training of Pos-Net. At inference time, only the fixed and moving image pair is required.

3.4. Network architectures

The proposed two-stage MUSA framework is compatible with most network architectures proposed for deep-learning-based DIR. We have selected some representative architectures to demonstrate the effectiveness and versatility of our method. We briefly introduce the selected architectures, with implementation detailed in Section 4.4.

3.4.1. Basic U-Net-based architectures

VoxelMorph (Balakrishnan et al., 2019) is selected as a U-Net representative. We also adopt a more recent residual U-Net design, referred to as Res-U-Net, based on the architecture described in the nnU-Net paper (Isensee et al., 2021), which has become a widely used benchmark for a variety of segmentation tasks.

3.4.2. Architectures with enlarged effective receptive fields (ERFs)

We select TransMorph (Chen et al., 2022) and LK-U-Net (Jia et al., 2022) to represent models with large effective receptive fields.

TransMorph is a hybrid Swin Transformer and CNN architecture, while LK-U-Net is pure CNN-based.

3.4.3. Architectures with explicit multiresolution modeling

All previously described architectures rely on implicit multi-resolution modeling of the neural network. Two architectures employing explicit multiresolution strategies have been selected: LapIRN (Mok and Chung, 2020) and Dual-PR-Net (Hu et al., 2019; Kang et al., 2022). They both predict the deformation on multiple scales and are optimized in a coarse-to-fine manner. However, LapIRN refines the deformation using an addition operator rather than a composition operator, which is not directly compatible with deformation composition in the MUSA framework. So, we only use it for the single-stage baseline. On the other hand, Dual-PR-Net uses a composition operator, and thus can be applied to all three schemes depicted in Fig. 1.

4. Experiments

4.1. Datasets and preprocessing

We evaluated the proposed MUSA framework for both inter-subject and intra-subject registration. Given the scarcity of paired intra-subject data, we reserved them exclusively for testing, whereas the inter-subject dataset was partitioned into training, validation, and test sets.

We built an inter-subject dataset from The Cancer Imaging Archive (TCIA) (Clark et al., 2013) from seven publicly available sources, including CPTAC-HNSC (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium, 2018), Head-Neck Cetuximab (Ang et al., 2014; Bosch et al., 2015), HEAD-NECK-RADIOMICS-HN1 (Aerts et al., 2014; Wee and Dekker, 2019), HNSCC (Elhalawani et al., 2017; Grossberg et al., 2020, 2018), HNSCC-3DCT-RT (Bejarano et al., 2019, 2018), QIN-HEADNECK (Beichel et al., 2015; Fedorov et al., 2016), and TCGA-HNSC (Zuley et al., 2016). Images were selected with the following criteria: 1) the field-of-view covers intact head-and-neck anatomy, 2) no foreign devices are present in the region of interest (typically above the C7 vertebra). A total of 380 images were selected and further partitioned into sets of 300, 40, and 40 images for training, validation, and testing. For training, the fixed and moving images were randomly paired from the training set. For validation and testing, inter-subject images were paired multiple times to expand the sample size. A total of 100 inter-subject pairs were generated randomly from the 40 images in validation or testing set while balancing the appearance of each image (i.e., each image appeared 5 times in the 100 pairs).

For the intra-subject dataset, an in-house dataset consisting of 7 patients was gathered for testing only. Each patient had one planning CT for radiation therapy and one PET attenuation correction CT from PET/CT. The latter was acquired in a nontreatment position as opposed to the immobilized planning CT. This results in large posture differences and large deformation between the two CT images, which represents the more challenging scenario of intra-subject head-and-neck CT registration (Hwang et al., 2009).

The same preprocessing pipeline was employed for both inter- and intra-subject datasets. To ensure the head-and-neck region was centered in the field-of-view, all images were rigidly registered to a pre-selected template using Elastix (Klein et al., 2010). Affine pre-registration is not needed since we use bending energy regularization, which allows linear deformation (Ding and Niethammer, 2022; Fischer and Modersitzki, 2003). Scanning beds and immobilization equipment were masked out of the images. Image intensity values were first clipped to a range of [-1024, 3000] Hounsfield Units (HU) and then normalized to the range of [0,1]. All volumes were resampled to an isotropic pixel spacing of $2 \times 2 \times 2$ mm using trilinear interpolation. Then, the volumes were cropped to the same matrix size of $160 \times 160 \times 192$. In two-stage multiresolution approach and two-stage MUSA approach, the first stage used half-resolution images ($4 \times 4 \times 4$ mm) of size $80 \times 80 \times 96$.

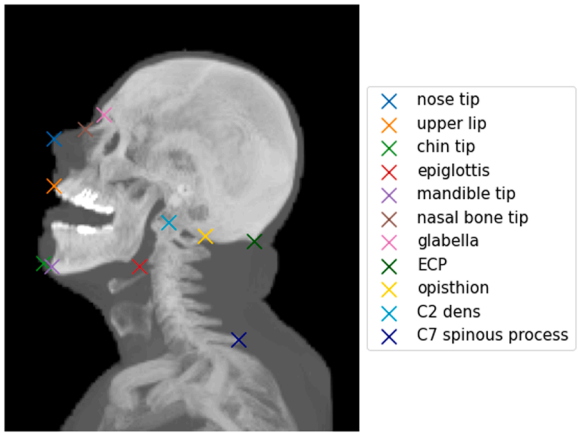

Fig. 3. An example of the 11 annotated landmarks used for target registration error (TRE) evaluation. The landmarks are overlaid on the CT image rendered using maximum intensity projection (MIP).

4.2. Evaluation metrics

Qualitative evaluation was performed by visually comparing the deformed images and the deformation vector fields across all methods. For quantitative evaluation, we used target registration error (TRE) to quantify landmark matching accuracy and three contour matching metrics, including Dice score, 95 percentile Hausdorff Distance (HD95 in mm), and Average Symmetric Surface Distance (ASD in mm). Further analysis of the deformation field was performed to evaluate deformation regularity and plausibility.

For TRE evaluation, 13 landmarks were labeled manually: the tip of the nose, the midpoint of the upper lip, the tip of the chin, the epiglottis, the lower front tip of the mandible, the tip of the nasal bone, the glabella, the external occipital protuberance (ECP), the opisthion, the left/right styloid process, the dens of the C2 vertebra, and the spinous process of the C7 vertebra. The landmarks were annotated by a medical physicist trained by a radiation oncologist specializing in head-and-neck cancer. All annotations were subsequently reviewed and approved by the same clinical expert to ensure accuracy and consistency. The left and right styloid processes were excluded from TRE calculation for hyper-parameter tuning and final evaluation as they were too small to be captured by the registration algorithm and resulted in excessively high errors across all registration methods, as shown in Fig. 5 (a). An example of the remaining 11 landmarks used for final evaluation is shown in Fig. 3. The landmarks were labeled on images with 1 mm resolution, and the final deformation field (2 mm resolution) was upsampled to 1 mm resolution via trilinear interpolation for TRE calculation. We averaged the TRE across the 11 landmarks for each registration pair. The mean and standard deviation of the average TRE across the test set were compared for different registration methods. For the contour-based metrics, we used 25 contours, including the brainstem, the cord, the chiasm, the left/right orbit, the left/right submandibular gland, the left/right optic nerve, the left/right parotid gland, the left/right cochlea, the larynx, the pharynx, the esophagus, the skull, the mandible, and cervical vertebrae C1-C7. The first 16 are common soft tissue organs at risk

(OARs) in head-and-neck radiation therapy, and the latter 9 are bony structures. The contours were automatically segmented using deep learning methods (Payer et al., 2020; Tong et al., 2019). All contours used for evaluation were reviewed and approved by the same clinical expert. Similar to TRE, the metrics, including Dice score, HD95, and ASD, were averaged for each registration pair, with the mean and standard deviation reported across the test set. For each architecture, we compared three approaches (a, b, and c) using the one-sided Wilcoxon signed-rank test to assess differences in quantitative metrics. We adjusted for multiple comparisons by applying the Bonferroni correction (Armstrong, 2014), dividing the significance level by the number of pairwise tests conducted (three per architecture).

We further employed several metrics to investigate the regularity and plausibility of deformation. The Jacobian determinant ($|J_\phi|$) of transformation ϕ estimates the local volume changes, where $|J_\phi| > 1$ indicates expansion, $0 < |J_\phi| < 1$ indicates shrinkage, and $|J_\phi| \leq 0$ indicates a singularity or folding. Following the deformation regularity assessment in Hering et al. (2023) , we calculated the percentages of non-positive values ($|J_\phi| \leq 0$, i.e., folded voxels) and the standard deviation of its logarithm ($SDlog|J_\phi|$). We also reported the median deformation magnitude and analyzed the deformation magnitude distribution to demonstrate the capability of each method to recover the large deformation. The Jacobian determinant metrics and median deformation magnitude results were averaged for each image pair and then aggregated for the entire inter- or intra-subject test set. In addition, we examined the Jacobian determinant maps to provide an intuitive assessment of deformation plausibility, which could reveal unrealistic expansion or shrinkage.

To better understand the reason for performance differences between different networks and different loss functions, we also analyzed the effective receptive fields (ERFs) of each method. Luo et al. (2016) introduced ERF to quantify the influence of each input voxel on the target output voxel. Previous studies have attributed the improved network performance to increased ERF both for CNNs (Ding et al., 2022; Jia et al., 2022) and Transformers (Chen et al., 2022; Li et al., 2023a). We adopted the ERF implementation in Ding et al. (2022). To calculate the ERF, a probing point (i, j, k) is selected in the output deformation vector $\boldsymbol{u}$. The probing point can be placed at the center of the field or within a specific region of interest, such as within a bony mask. The deformation vector at the probing point $\boldsymbol{u}(i, j, k)$ is then

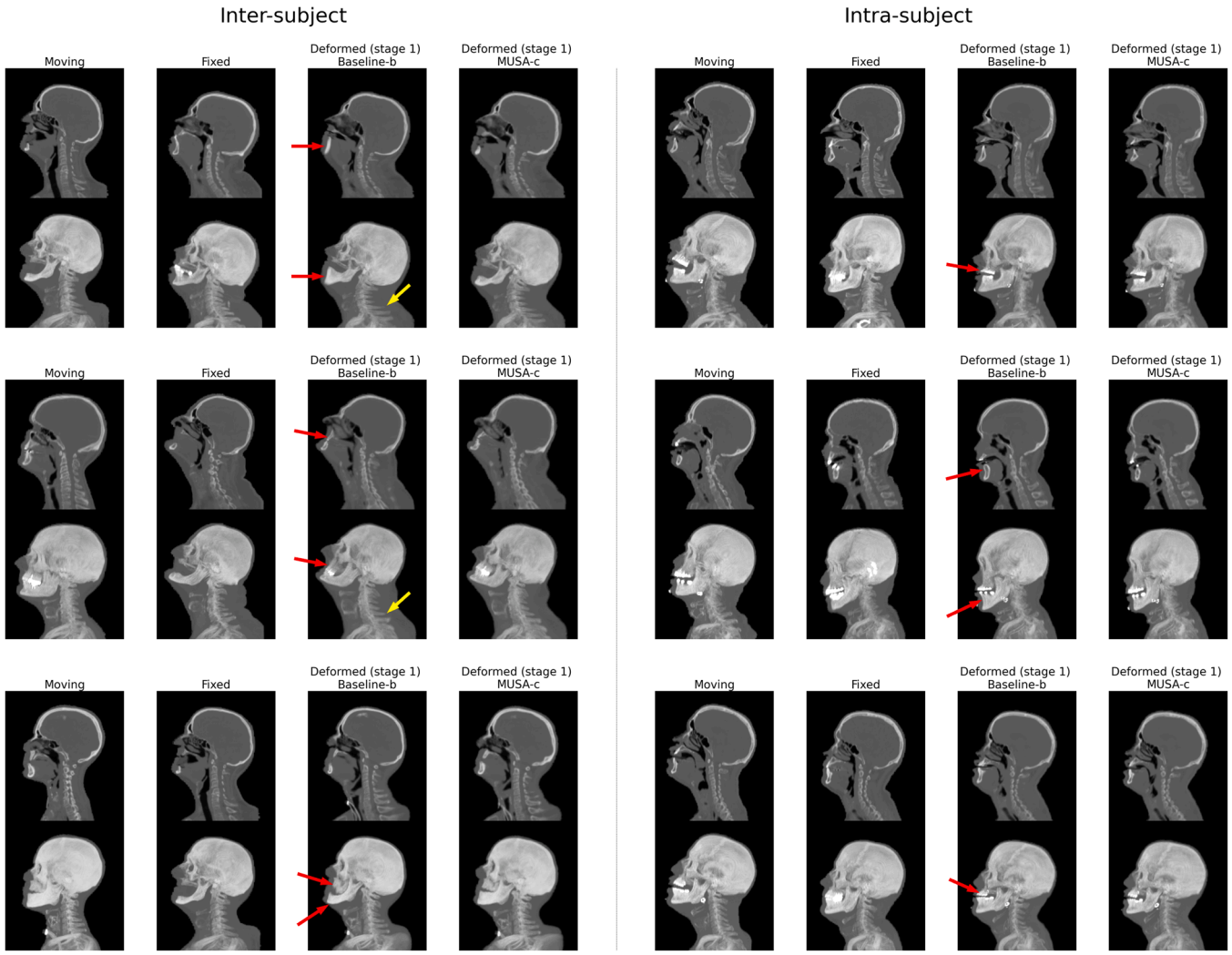

Fig. 4. Posture correction results obtained from Pos-Net using the proposed MUSA loss. The left and right panels demonstrate results for three example cases of inter- and intra-subject registration, respectively. For each case, the moving image, the fixed image, the first-stage deformed image of the two-stage multiresolution baseline (b), and the first-stage deformed image of the MUSA approach (c) are displayed. Both the central slice and the maximum intensity projection (MIP) rendering from the coronal view are presented. Arrows in the deformed images highlight the differences between the two approaches. The TransMorph architecture was used for all results shown.

backpropagated to the input fixed and moving images. The resulting gradient indicates how much each voxel of the fixed or moving image contributes to the final deformation vector at (i, j, k). This reflects where the network is attending to when predicting the deformation at the probing location. L2 normalization was applied to compare ERFs across different networks.

4.3. MUSA configuration

This section elaborates on detailed settings related to the MUSA loss and MUSA registration framework, including mask preprocessing, resolution selection and hyperparameter tuning.

The bony masks used to calculate MUSA loss in Eq. (5) underwent preprocessing with morphological erosion on the touching boundaries of vertebra and mandible. By doing so, we ensured the proper separation of all bony structures from one another to prevent overt coupling in motion estimation, a known issue with the rigidity constraint in Staring et al. (2007) (Kim et al., 2013).

The Pos-Net with MUSA loss can be trained on full- or half-resolution images. We found that training on half-resolution images was not only more efficient, but also yielded better and more stable posture correction results, as the bulk posture change primarily involved global coarse deformation. Among all the network architectures evaluated in this study, only TransMorph and Dual-PR-Net successfully sustained stable training using full-resolution images, while all networks demonstrated reasonable performance when trained on half-resolution images. Further reduction in the resolution (e.g., $8 \times 8 \times 8$ mm) proved to be impractical, as the bone segmentations became excessively coarse.

The regularization hyperparameters (λ and α) in MUSA loss (i.e., Eq. (5)) balance the trade-off between intensity matching and deformation regularity. Although a two-dimensional grid search for λ and α across all network architectures would optimize the performance, it is computationally impractical. In the posture correction stage, λ controls the deformation regularity for soft tissue, which makes up most of the body. In the refinement stage, λ controls the global deformation regularity. The optimal λ should align closely with that of the two-stage multiresolution approach (baseline (b)). Hence, we set λ for MUSA loss to match the optimal λ for baseline (b). The optimal λ for single-stage and two-stage baselines was tuned independently for each architecture considering both Dice and TRE. The detailed tuning process and results are described in Appendix A1. Then, we fixed λ and tuned α only for MUSA loss. The tuning was performed on three representative architectures with varying capacity, including VoxelMorph, TransMorph, and Dual-PR-Net. The optimal α value was determined using TRE on the validation set. The tuning process and results are detailed in Appendix A2. We observed that all models achieved the best TRE when α equaled 1000. The degree of improvement was, however, dependent on model capacity: VoxelMorph saw the most improvement, TransMorph showed moderate improvement, while Dual-PR-Net showed negligible gains due to its powerful explicit multiresolution modeling. The consistent results across different models, despite varying capacities, can be attributed to the inherent physical meaning of α, which characterizes the relative rigidity of bony structures compared to soft tissue. The Dice scores were not sensitive to α. Introducing MUSA loss only marginally improved the observed bony Dice without affecting the Dice for unobserved soft tissue organs. As a result, α of 1000 was universally used for all architectures. It also demonstrated good qualitative posture correction performance, as later shown in Fig. 4. The ERF analysis, discussed in Section 5.2.4, also justified the selection of $\alpha = 1000$.

Table 1
The number of parameters in each network architecture. The values are in units of millions of parameters. These counts are for a single network. For two-stage registration, the total number doubles.

Model	Parameters (M)
VoxelMorph	0.33
Res-U-Net	16.67
LK-U-Net	7.81
TransMorph	46.75
Dual-PR-Net	0.49
LapIRN	0.92

4.4. Implementation details for the proposed and baseline methods

We adopted B-spline registration using Elastix (Klein et al., 2010) as the conventional iterative baseline. It integrates the explicit multiresolution strategy with Gaussian pyramid. The parameters were empirically optimized and demonstrated good performance on head-and-neck CT registration (McKenzie et al., 2020).

For deep-learning-based methods, six network architectures described in Section 3.4 were utilized: VoxelMorph, Res-U-Net, LK-U-Net, TransMorph, Dual-PR-Net, and LapIRN. To demonstrate the effectiveness of our proposed MUSA framework, we trained these networks under three different configurations: (a) single-stage baseline, (b) two-stage multiresolution baseline, and (c) two-stage MUSA approach, as depicted in Fig. 1. LapIRN was only tested with single-stage baseline, as it already had the multiresolution setup and its deformation refinement via addition was different from the refinement via composition used in our multiresolution scheme.

During each training epoch, 100 image pairs were randomly generated from the training set. The single-stage framework was trained for 1500 epochs on full-resolution images. The two-stage frameworks can be trained sequentially: initially training the first network, and then appending the second network. It is also feasible to initialize the second network with trained weights from the single-stage models. We found this method to be more efficient, with no detriment to the final performance. Therefore, for all two-stage approaches, we initially trained the first and second-stage models on half and full-resolution images, respectively, for 1000 epochs. Subsequently, the two models were concatenated and trained for an additional 500 epochs. During this combined training phase, the weights of the first-stage model were fixed, allowing only the second-stage model to update.

All the methods were implemented using PyTorch (Paszke et al., 2019) on the same system equipped with an Intel(R) Core(TM) i9-10900X CPU and NVIDIA Quadro RTX 8000 GPU. All architectures employed their non-diffeomorphic versions, given that we were training on an inter-subject dataset exhibiting substantial anatomical variances among the images. The implementation details for different architectures were as follows.

- VoxelMorph[1] (Balakrishnan et al., 2019): We employed the official implementation of VoxelMorph with default settings.
- Res-U-Net[2] (Isensee et al., 2021): We replaced the 3D U-Net in VoxelMorph with a more recent 3D residual U-Net implementation from the nnU-Net paper (Isensee et al., 2021). The other parts were kept the same as in VoxelMorph.
- LK-U-Net[3] (Jia et al., 2022): We experimented with different large kernel sizes (5/7/9/11/13/15). Kernel size 9 yielded the best result on the validation set and was used for all experiments. All other implementations were consistent with the authors' implementation.
- TransMorph[4] (Chen et al., 2022): We employed the official implementation of TransMorph with the default settings. The image size and window size were (160,160,192) and (5,6,6) for full-resolution training and (128,128,128) and (4,4,4) for half-resolution training.

[1] https://github.com/voxelmorph/voxelmorph.
[2] https://github.com/MIC-DKFZ/nnUNet.
[3] https://github.com/xi-jia/LKU-Net.
[4] https://github.com/junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration.

Table 2
Average inference time for methods compared in this study. Elastix used CPUs, while the deep learning methods used GPU. The average inference time was calculated based on 100 repeated runs. The forward passes of the two-stage multiresolution and the two-stage MUSA approach are identical, therefore they have the same inference time.

Method	Single-stage (sec/pair)	Two-stage (sec/pair)
Elastix	37.9	-
VoxelMorph	0.194	0.227
Res-U-Net	0.318	0.375
LK-U-Net	0.176	0.211
TransMorph	0.255	0.362
Dual-PR-Net	0.229	0.266
LapIRN	0.223	-

Zero-padding was used for half-resolution images to ensure that the image size was divisible by the window size.

- Dual-PR-Net[5] (Hu et al., 2019; Kang et al., 2022): We implemented the network ourselves following the description provided in the papers. Our implementation used the residual convolutions proposed in Dual-PR-Net++ (Kang et al., 2022) but omitted the 3D correlation layer due to its high computational demands. We observed that adding the residual convolutions alone was sufficient to achieve top performances in head-and-neck CT registration.
- LapIRN[6] (Mok and Chung, 2020): We adhered to the official LapIRN implementation using a three-level pyramid and sequential training strategies. The three levels were trained for 300, 300, and 900 epochs, respectively, resulting in a total of 1500 epochs.

Table 1 lists the number of trainable parameters for each network architecture. Table 2 compares the inference time of all methods studied in this paper. For deep-learning-based methods, the two-stage framework led to a modest increase in inference time, ranging between 16-20%, with TransMorph as an exception at 42% due to increased input size from padding, when compared to the single-stage approach.

4.5. Comparison with Dice loss

The bony segmentations in MUSA loss are used to encourage deformation regularity in contrast to enforcing contour matching as in Dice loss, proposed in Balakrishnan et al. (2019). To better understand this difference and justify the contribution of MUSA loss, we performed an addition experiment with the Dice loss baseline. Its loss function is defined as:

$$Loss = MSE(f, m\circ\phi) + \lambda\sum_{\boldsymbol{r}\in\Omega_f} BE(\phi(\boldsymbol{r})) + \gamma L_{seg}\left(s_f, s_m\circ\phi\right), \tag{6}$$

where s_f and s_m are the segmentation masks of the fixed and moving image, respectively, and γ is the weighting parameter for L_{seg}. L_{seg} is calculated using Dice scores of segmentation masks:

$$L_{seg}\left(s_f, s_m\circ\phi\right) = 1 - \frac{1}{K}\sum_{k=1}^{K} Dice\left(s_f^k, s_m^k\circ\phi\right). \tag{7}$$

To ensure a fair comparison against MUSA loss, only bony segmentations were used in the Dice loss. The networks were trained with the same strategy as the two-stage multiresolution baseline (i.e., baseline (b)), with only the loss function modified. The hyperparameter tuning process for γ is detailed in Appendix B1. The final experiment was carried out with the TransMorph architecture as a representative, using the optimally tuned γ of 10^{-4}.

[5] https://github.com/kangmiao15/Dual-Stream-PRNet-Plus.
[6] https://github.com/cwmok/LapIRN.

5. Results

5.1. Posture correction results

We first demonstrated the posture correction performance of Pos-Net by comparing the first-stage results between MUSA and the two-stage multiresolution baseline. The qualitative results of both inter- and intra-subject registration using the TransMorph architecture are presented in Fig. 4. For the two-stage multiresolution baseline, the first-stage network achieved reasonable intensity matching, however, there were excessive deformations of bony structures, most evident in the upper and lower jaw regions and occasionally in the spine areas, as indicated by arrows in Fig. 4. This was more obvious in the inter-subject registration, due to large variations among patient anatomies. Slight implausible warping of bones was also observed in intra-subject cases. In contrast, Pos-Net successfully adjusted the posture of the moving image to more closely resemble that of the fixed image while ensuring that the bony structures were not deformed unrealistically. In this way, it focused on addressing the musculoskeletal motion, including head pitching, neck flexion and extension, and jaw movement, while the residual deformation was reserved for Ref-Net in the second stage. Although the intensity matching after Pos-Net was compromised compared to the two-stage multiresolution baseline, the posture correction step can assist in avoiding local minima and provide a better intermediate alignment that eases the burden of subsequent fine-scale registration. Fig. 4 also illustrates the limitations of both the baseline and proposed method in handling topological changes caused by mouth closing and opening (the second inter-subject example and the first and the third intra-subject example), which require specialized techniques.

Table 3
The average target registration error (TRE in mm) results for different methods on inter- and intra-subject test sets. For deep-learning-based methods, the three columns from left to right are (a) single-stage registration, (b) two-stage multi-resolution registration, and (c) two-stage MUSA registration, respectively. Methods employing explicit multiresolution modeling are indicated by an underline. When comparing columns a, b, and c, the best result within one row is highlighted in **bold**. The superscripts denote statistical significance. The single superscript in column (b) indicates statistical significance when comparing to baseline (a), i.e., b vs a. The two superscripts in column (c) indicate statistical significance when comparing to baseline (a) and (b), i.e., c vs a and c vs b, respectively. Superscripts [ns, s1, s2, s3, s4] correspond to statistical significance levels of α = [no significance, 0.05/3, 0.01/3, 0.001/3, 0.0001/3].

Method	Inter-subject test set (N=100)			Intra-subject test set (N=7)		
	a	b	c	a	b	c
Initial	18.69 ± 7.35	-	-	24.80 ± 8.04	-	-
Elastix	7.99 ± 3.73	-	-	3.92 ± 0.71	-	-
VoxelMorph	8.78 ± 4.12	6.46 ± 2.83 ^{s4}	**5.48 ± 1.95** ^{s4,s4}	11.26 ± 5.52	6.41 ± 3.39 ^{s1}	**3.81 ± 1.27** ^{s1,s1}
Res-U-Net	6.00 ± 2.46	5.18 ± 2.06 ^{s4}	**4.73 ± 1.62** ^{s4,s4}	6.86 ± 3.29	4.29 ± 2.20 ^{s1}	**2.93 ± 0.63** ^{s1,s1}
LK-U-Net	7.20 ± 2.47	5.60 ± 2.01 ^{s4}	**5.29 ± 1.76** ^{s4,s4}	8.55 ± 3.71	4.15 ± 1.49 ^{s1}	**3.30 ± 0.85** ^{s1,s1}
TransMorph	6.62 ± 2.63	5.04 ± 1.76 ^{s4}	**4.61 ± 1.32** ^{s4,s4}	6.38 ± 2.74	3.26 ± 0.85 ^{s1}	**2.64 ± 0.34** ^{s1,s1}
Dual-PR-Net	5.00 ± 1.51	4.91 ± 1.49 ^{ns}	**4.62 ± 1.23** ^{s4,s4}	3.56 ± 0.43	3.03 ± 0.62 ^{ns}	**2.69 ± 0.37** ^{s1,ns}
LapIRN	5.76 ± 2.49	-	-	3.35 ± 0.86	-	-

5.2. Registration results

5.2.1. Quantitative results

The average target registration error (TRE) comparison across all the frameworks and architectures is provided in Table 3. The initial TRE was large in both inter-subject (18.7 mm) and intra-subject (24.8 mm) datasets. The non-DL baseline Elastix performed well on the intra-subject test set, achieving TRE slightly worse than the best-performing DL-based methods. However, its performance declined sharply for inter-subject registration. For DL-based architectures, we observed an increase in model capacity based on their registration performance in the following order: VoxelMorph $<$ LK-U-Net $<$ Res-U-Net $\approx$ TransMorph $<$ Dual-PR-Net $\approx$ LapIRN. For the single-stage approach, only models incorporating explicit multiresolution modeling (i.e., Dual-PR-Net and LapIRN) achieved reasonable TRE results, while all other architectures (i.e., VoxelMorph, Res-U-Net, LK-U-Net, and TransMorph) had considerably higher TRE. The two-stage multiresolution approach and the two-stage MUSA approach reduced the TRE, with the two-stage MUSA approach consistently delivering the best results across all architectures. Consequently, we focus on the comparison between the two-stage MUSA approach (c) and the two-stage multiresolution baseline (b). The MUSA approach consistently improved TRE compared to baseline (b) across all five tested architectures on both datasets. The effect size (i.e., the difference in mean relative to standard deviation) decreased with increased model capacity, with VoxelMorph showing the largest improvement and Dual-PR-Net exhibiting only slight improvement. Statistical testing with Bonferroni correction confirmed that the TRE improvements on the inter-subject dataset were statistically significant, with very small p-values ($p<0.0001/3$) for all five architectures. For the intra-subject dataset, statistically significance ($p<0.05/3$) was observed for the four architectures without explicit multiresolution modeling, while Dual-PR-Net showed no significance ($p>0.05/3$). However, the effect size was larger on the intra-subject dataset than inter-subject results. Fig. 5 illustrates the performance with respect to each landmark. The left and right styloid processes had excessively high TRE across all methods in the inter-subject dataset, as these structures were too fine to be captured by the registration. Therefore, they were excluded from the average TRE calculation to avoid bias.

The contour-based metrics for all the methods are presented in Table 4, including the Dice score, HD95, and ASD. Dual-PR-Net and LapIRN once again achieved better performance than other architectures for the single-stage registration. When either of the two-stage framework was used, the results for architectures without explicit multiresolution modeling improved significantly. The proposed MUSA

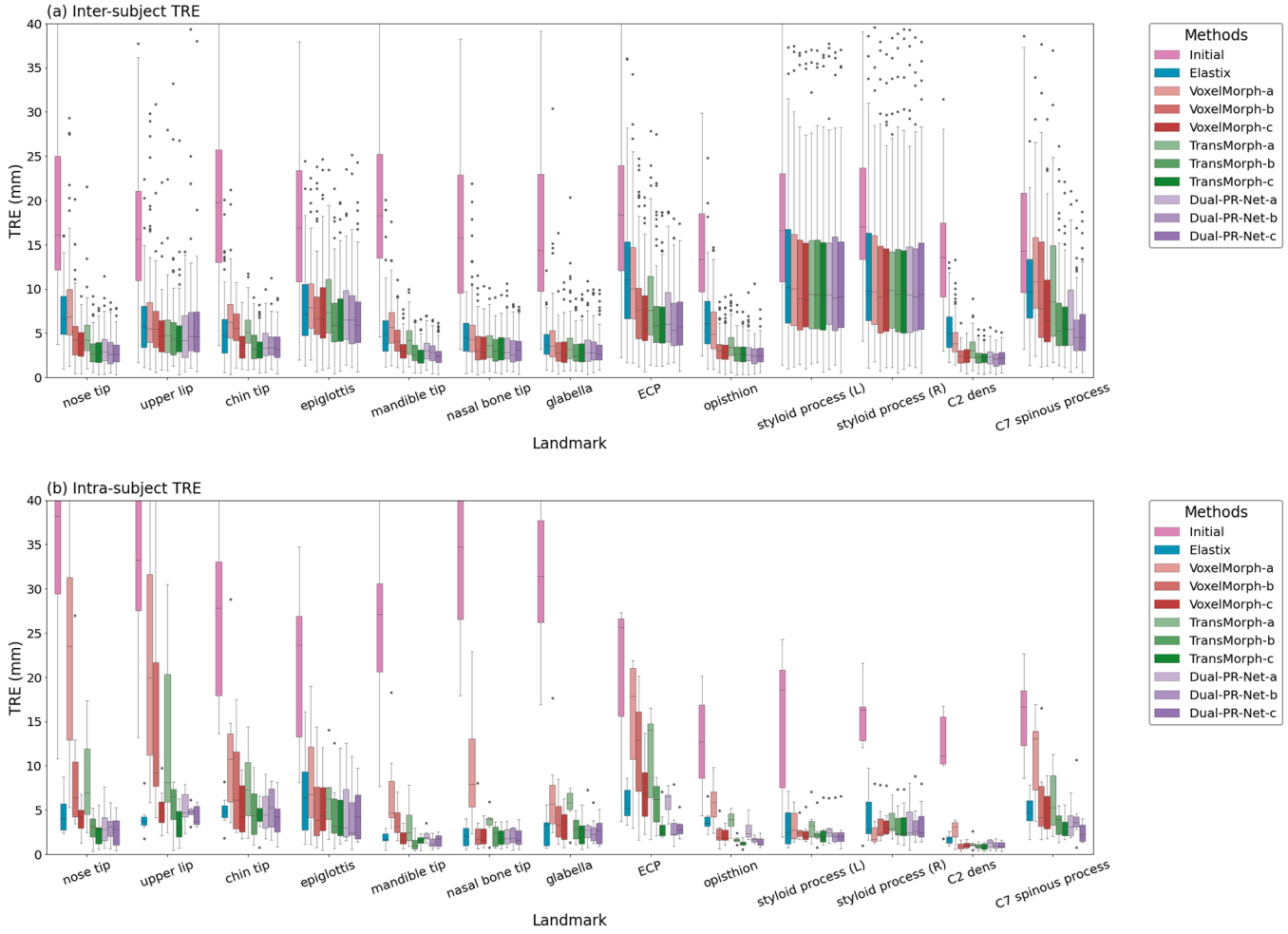

Fig. 5. Boxplots of the target registration error (TRE) for different landmarks in inter-subject (5a) and intra-subject (5b) registration. Due to space limitations, three representative network architectures are shown: VoxelMorph, TransMorph, and Dual-PR-Net. The letter suffixes a/b/c in the legend denote three comparison configurations: (a) single-stage registration, (b) two-stage multiresolution registration, and (c) two-stage MUSA registration, respectively. The left and right styloid processes had excessively high TRE across all methods in inter-subject registration, as these structures were too fine to be captured by the registration. Therefore, they were excluded from the average TRE calculation to avoid bias.

Table 4
Contour-based metrics for different methods on inter- and intra-subject test sets including Dice score (Dice), 95 percentile Hausdorff Distance (HD95 in mm), and Average Symmetric Surface Distance (ASD in mm). The table organization and notations are identical to Table 3.

Dice ↑						
Method	Inter-subject test set (N=100)			Intra-subject test set (N=7)		
	a	b	c	a	b	c
Initial	0.250 ± 0.094	-	-	0.209 ± 0.100	-	-
Elastix	0.527 ± 0.074	-	-	0.713 ± 0.040	-	-
VoxelMorph	0.542 ± 0.063	0.661 ± 0.055 ^{s4}	**0.668 ± 0.046** ^{s4,ns}	0.619 ± 0.078	0.729 ± 0.046 ^{s1}	**0.742 ± 0.035** ^{s1,ns}
Res-U-Net	0.650 ± 0.044	0.700 ± 0.040 ^{s4}	**0.705 ± 0.034** ^{s4,s1}	0.715 ± 0.050	0.748 ± 0.039 ^{s1}	**0.755 ± 0.032** ^{s1,ns}
LK-U-Net	0.589 ± 0.047	0.677 ± 0.042 ^{s4}	**0.680 ± 0.034** ^{s4,ns}	0.663 ± 0.050	0.739 ± 0.037 ^{s1}	**0.743 ± 0.034** ^{s1,ns}
TransMorph	0.638 ± 0.047	0.698 ± 0.041 ^{s4}	**0.703 ± 0.037** ^{s4,ns}	0.703 ± 0.047	0.754 ± 0.032 ^{s1}	**0.755 ± 0.029** ^{s1,ns}
Dual-PR-Net	0.682 ± 0.043	0.692 ± 0.045 ^{s2}	**0.707 ± 0.030** ^{s4,s1}	0.738 ± 0.029	0.753 ± 0.031 ^{ns}	**0.755 ± 0.032** ^{s1,ns}
LapIRN	0.662 ± 0.064	-	-	0.757 ± 0.034	-	-
HD95 ↓						
Method	Inter-subject test set (N=100)			Intra-subject test set (N=7)		
	a	b	c	a	b	c
Initial	16.07 ± 4.52	-	-	16.89 ± 4.37	-	-
Elastix	8.72 ± 2.03	-	-	5.41 ± 1.21	-	-
VoxelMorph	8.34 ± 1.74	6.84 ± 1.51 ^{s4}	**6.55 ± 1.26** ^{s4,s2}	6.67 ± 1.67	5.33 ± 1.20 ^{s1}	**4.80 ± 1.03** ^{s1,s1}
Res-U-Net	6.69 ± 1.33	6.16 ± 1.33 ^{s4}	**6.02 ± 1.22** ^{s4,s2}	5.37 ± 1.29	4.95 ± 1.24 ^{s1}	**4.67 ± 0.94** ^{s1,ns}
LK-U-Net	7.40 ± 1.21	6.52 ± 1.36 ^{s4}	**6.35 ± 1.23** ^{s4,s2}	6.16 ± 1.24	4.97 ± 1.09 ^{s1}	**4.86 ± 1.02** ^{s1,ns}
TransMorph	6.94 ± 1.31	6.18 ± 1.28 ^{s4}	**6.00 ± 1.23** ^{s4,s3}	5.32 ± 0.97	4.64 ± 0.96 ^{s1}	**4.61 ± 0.90** ^{s1,ns}
Dual-PR-Net	6.43 ± 1.45	6.39 ± 1.27 ^{ns}	**6.04 ± 1.19** ^{s4,s3}	4.91 ± 0.90	4.82 ± 0.95 ^{ns}	**4.65 ± 0.92** ^{s1,ns}
LapIRN	7.01 ± 1.84	-	-	4.71 ± 0.99	-	-
ASD↓						
Method	Inter-subject test set (N=100)			Intra-subject test set (N=7)		
	a	b	c	a	b	c
Initial	7.96 ± 3.00	-	-	8.50 ± 2.69	-	-
Elastix	3.25 ± 1.07	-	-	1.70 ± 0.31	-	-
VoxelMorph	3.08 ± 0.89	2.32 ± 0.80 ^{s4}	**2.23 ± 0.75** ^{s4,s1}	2.25 ± 0.55	1.62 ± 0.29 ^{s1}	**1.50 ± 0.24** ^{s1,s1}
Res-U-Net	2.31 ± 0.80	2.04 ± 0.77 ^{s4}	**2.01 ± 0.75** ^{s4,ns}	1.69 ± 0.36	1.50 ± 0.28 ^{s1}	**1.43 ± 0.22** ^{s1,ns}
LK-U-Net	2.67 ± 0.75	2.18 ± 0.77 ^{s4}	**2.14 ± 0.77** ^{s4,ns}	2.02 ± 0.38	1.54 ± 0.25 ^{s1}	**1.51 ± 0.23** ^{s1,ns}
TransMorph	2.39 ± 0.76	2.05 ± 0.75 ^{s4}	**2.01 ± 0.78** ^{s4,ns}	1.72 ± 0.26	1.43 ± 0.22 ^{s1}	**1.42 ± 0.20** ^{s1,ns}
Dual-PR-Net	2.15 ± 0.78	2.12 ± 0.80 ^{ns}	**1.99 ± 0.76** ^{s4,s1}	1.53 ± 0.16	1.46 ± 0.22 ^{ns}	**1.43 ± 0.22** ^{s1,ns}
LapIRN	2.39 ± 0.95	-	-	1.43 ± 0.23	-	-

approach consistently achieved the best results across all architectures. However, the improvement of Dice score and ASD was marginal and there was no stable statistical significance across different architectures. Interestingly, for HD95, we observed statistically significant improvement in the inter-subject dataset for all five architectures. Since HD95 measures maximum discrepancies in contour alignment (excluding the most extreme 5%), this result suggested that MUSA could reduce the outlier errors in contour alignment, providing a more robust registration. Fig. 6 shows Dice scores for different anatomical structures.

5.2.2. *Qualitative results*

Fig. 7 presents the qualitative results of two representative cases from inter-subject registration and intra-subject registration, respectively. The deformed images and their corresponding deformation fields, visualized as quiver and grid plots, are displayed for all compared frameworks and architectures. All four single-stage methods using architectures without explicit multiresolution modeling (i.e., VoxelMorph, Res-U-Net, LK-U-Net, and TransMorph) exhibited obvious registration errors in the mandible, chin, and nose areas, revealing their limitations in handling large deformations. The severity of these errors decreased as the model capacity increased. In contrast, Dual-PR-Net and LapIRN produced more plausible deformed images. Elastix performed well in intra-subject registration but degraded in inter-subject registration as the B-spline registration exhibited excessive smoothness, making it challenging to address complex deformation. The two-stage multiresolution approach markedly improved the deformed images for architectures lacking explicit multiresolution modeling, however, some errors were still visible for architectures with low registration capacity (e.g., VoxelMorph). The two-stage MUSA approach further improved the registration, and the deformed images were more visually consistent across architectures with varying capacity. The quiver and grid plots further demonstrated that the MUSA approach generated the most consistent deformation fields across different architectures compared to the two baselines.

The quiver plots in Fig. 7 also revealed differences in how well each method captured the global head rotation (pitch movement), a major posture difference between the fixed and moving images. For the single-stage framework, all four architectures without explicit multiresolution modeling failed to capture the head rotation, as indicated by the absence of deformation in the upper rear part of the head (highlighted by the dotted ellipse in the quiver plots). The two-stage multiresolution approach also failed to capture such deformation. On the other hand, models with explicit multiresolution modeling (Dual-PR-Net and LapIRN) better reflected the global posture change. Again, Elastix performed well for the intra-subject case but fell short for the inter-subject case. Our proposed two-stage MUSA approach effectively captured the head rotation across all network architectures, although there were still underestimations in the magnitude in architectures with lower capacity (e.g., VoxelMorph and LK-U-Net). These results demonstrate that the MUSA approach can better represent postural differences in head-and-

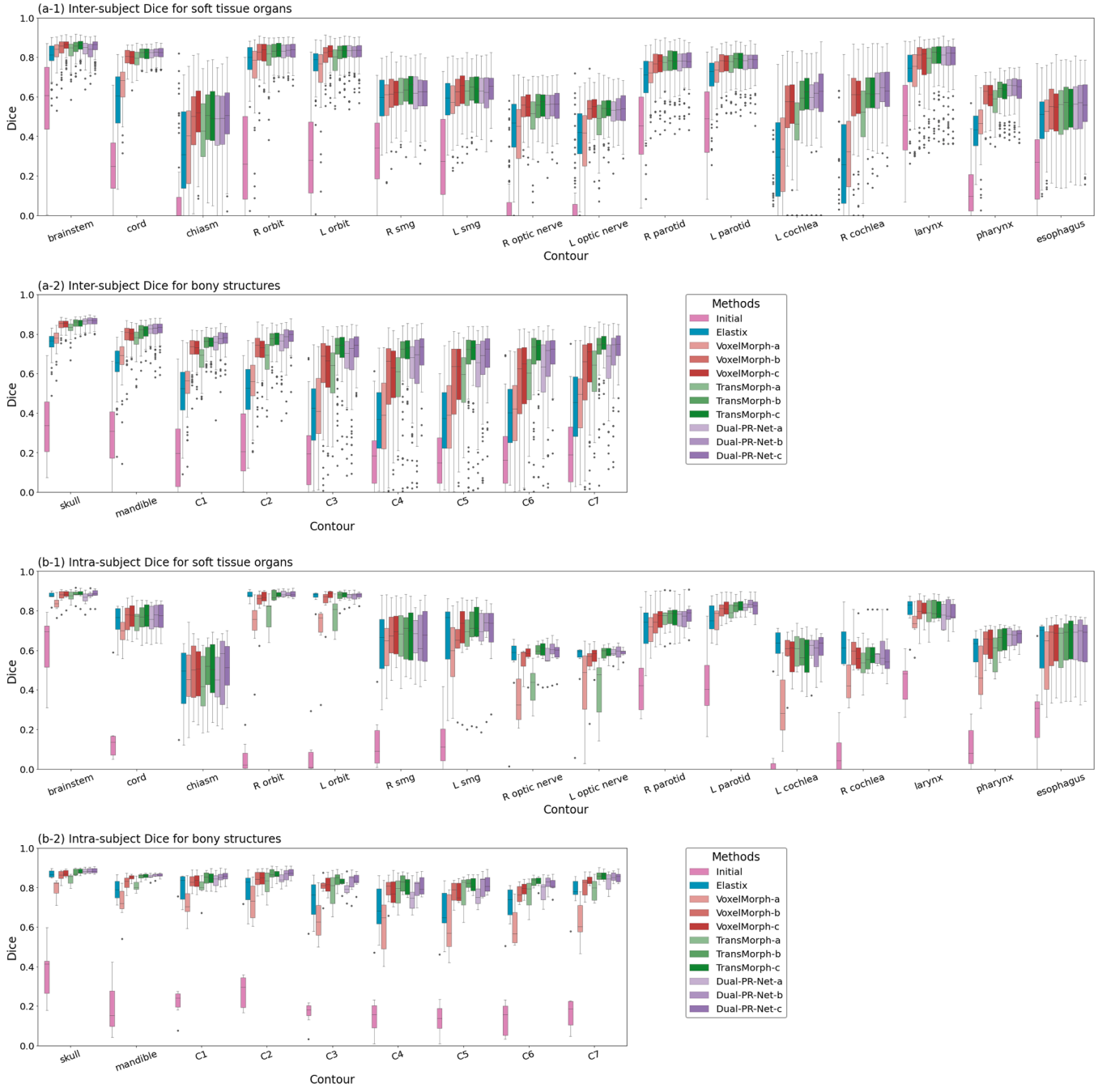

Fig. 6. Boxplots of Dice scores for different anatomical structures. Inter-subject results are shown in (6a-1/6a-2) and intra-subject results in (6b-1/6b-2). Soft tissue organs (6a-1/6b-1) and bony structures (6a-2/6b-2) are displayed in separate plots. Due to space limitations, three representative network architectures are shown: VoxelMorph, TransMorph, and Dual-PR-Net. The letter suffixes a/b/c in the legend denote three comparison configurations: (a) single-stage registration, (b) two-stage multiresolution registration, and (c) two-stage MUSA registration, respectively.

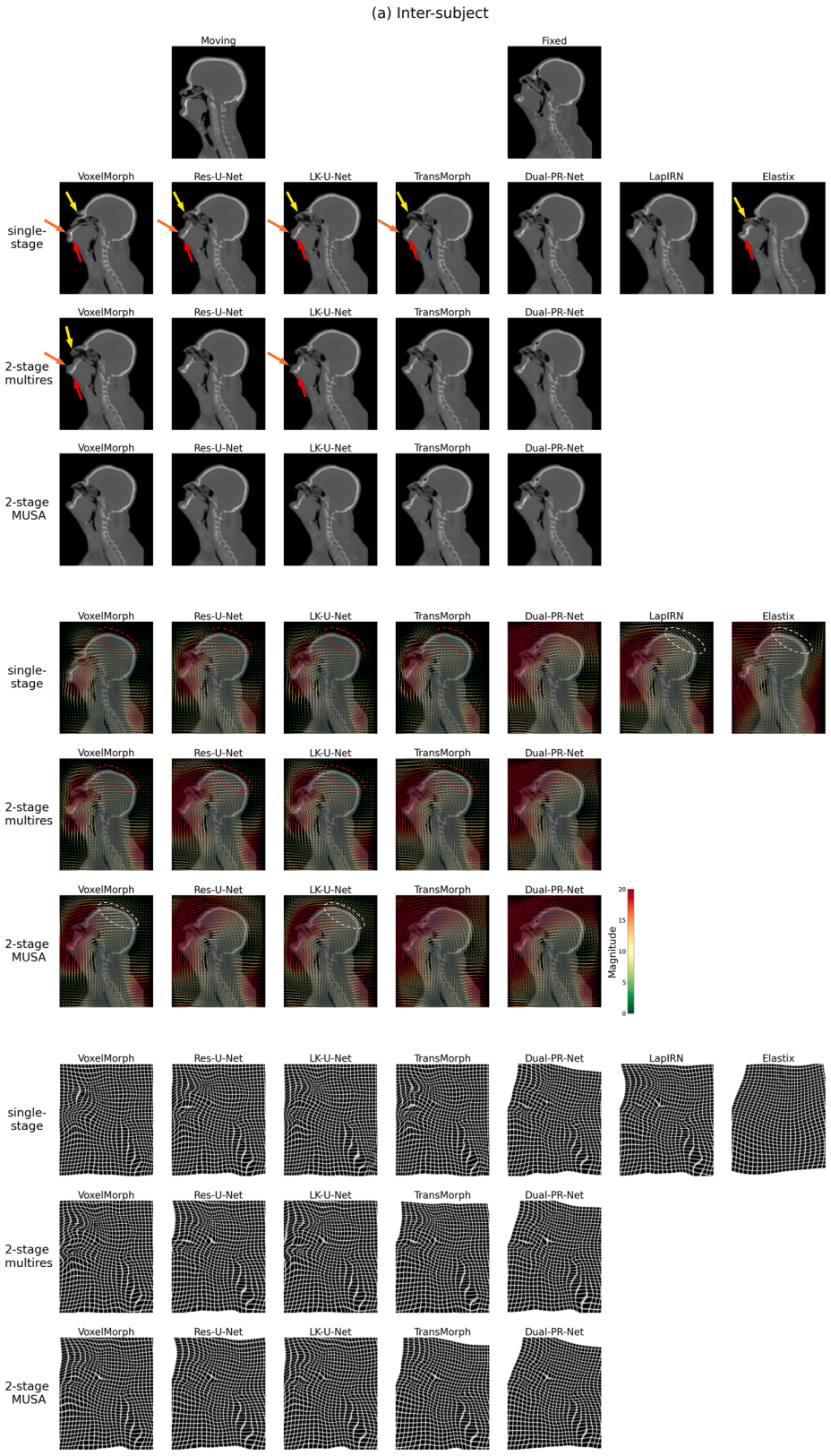

Fig. 7. Qualitative results of (7a) inter-subject and (7b) intra-subject registration. The two-stage MUSA registration results are compared to the single-stage baseline and the two-stage multiresolution baseline. The central slice of the coronal view is displayed due to the presence of most deformation in the superior-to-inferior (SI) and anterior-to-posterior (AP) directions. The moving and fixed images are shown at the top, followed by the deformed images, quiver plots of the deformation field, and grid plots of the deformation field. Arrows in the deformed images highlight the registration errors. The dotted ellipses in the quiver plots highlight the regions of interest where discrepancies in head rotation (pitch) are evident. Specifically, red ellipses indicate failure to recover the rotational motion, while white ellipses indicate underestimation of the motion magnitude. The last three models (i.e., Dual-PR-Net, LapIRN and Elastix) has explicit multiresolution modeling.

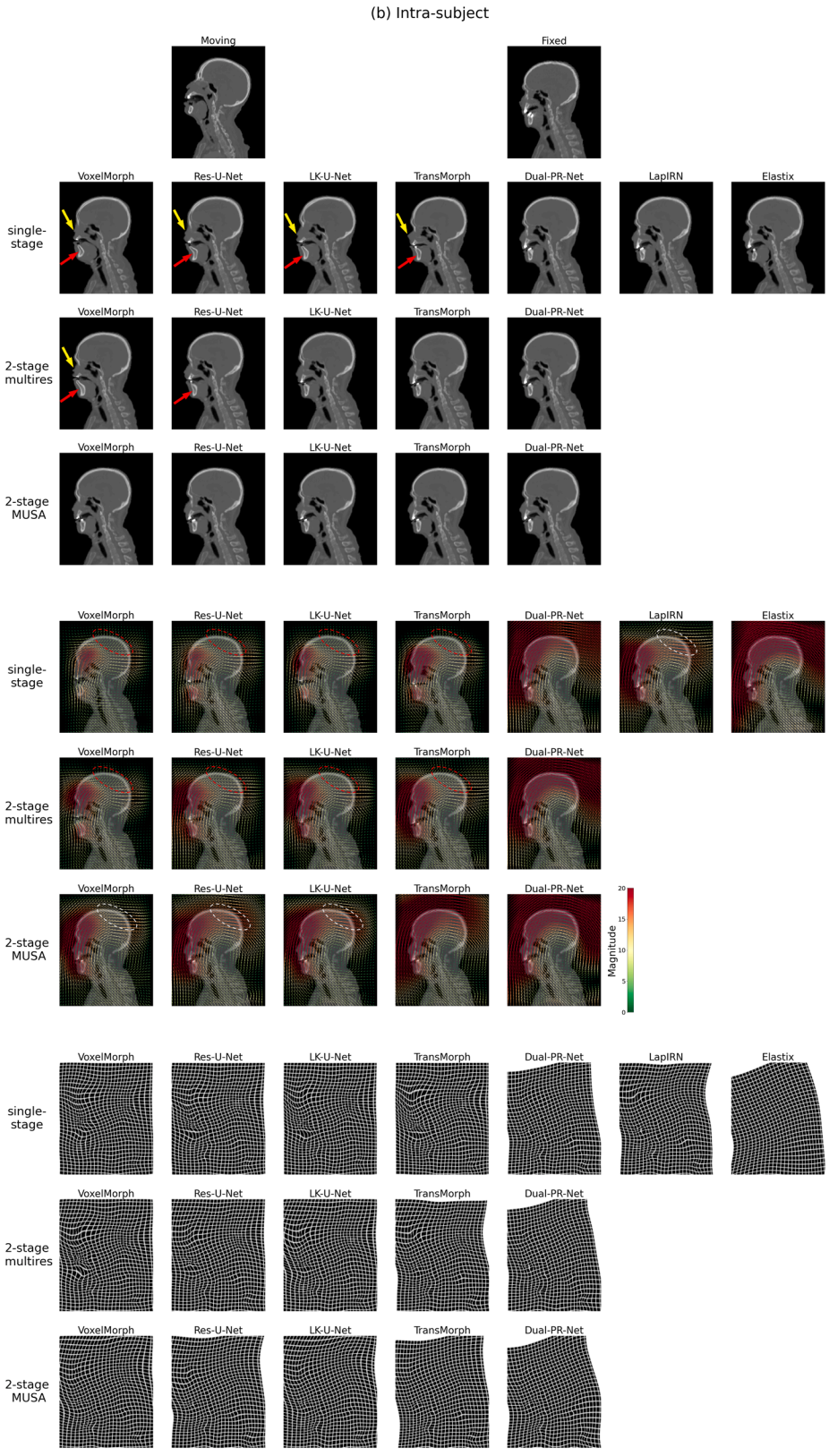

Fig. 7. (*continued*).

Table 5
Deformation-based metrics of different methods on inter- and intra-subject test sets. We reported the percentage of folded voxels (% of $|J_\phi| \leq 0$), the standard deviation of the logarithm of the Jacobian determinant ($SDlog|J_\phi|$), and the median magnitude of deformation vectors within the body mask.

% of $\|J_\phi\| \leq 0$						
Method	Inter-subject test set (N=100)			Intra-subject test set (N=7)		
	a	b	c	a	b	c
Elastix	0.0000 ± 0.0000	-	-	0.0000 ± 0.0000	-	-
VoxelMorph	0.0057 ± 0.0376	0.0104 ± 0.0637	0.0077 ± 0.0346	0.0004 ± 0.0010	0.1932 ± 0.3248	0.0785 ± 0.0951
Res-U-Net	0.0302 ± 0.1122	0.0493 ± 0.2028	0.0151 ± 0.0392	0.3055 ± 0.3637	0.4166 ± 0.4906	0.2904 ± 0.3541
LK-U-Net	0.0577 ± 0.2746	0.0080 ± 0.0403	0.0181 ± 0.0806	0.2831 ± 0.3296	0.2617 ± 0.3016	0.3097 ± 0.3659
TransMorph	0.0468 ± 0.1878	0.0384 ± 0.1566	0.0312 ± 0.1307	0.3413 ± 0.4191	0.0279 ± 0.0490	0.0154 ± 0.0244
Dual-PR-Net	0.0013 ± 0.0046	0.0051 ± 0.0244	0.0004 ± 0.0029	0.1321 ± 0.2006	0.0000 ± 0.0000	0.0102 ± 0.0218
LapIRN	0.1021 ± 0.2031	-	-	0.5122 ± 0.6481	-	-
$SDlog\|J_\phi\|$						
Method	Inter-subject test set (N=100)			Intra-subject test set (N=7)		
	a	b	c	a	b	c
Elastix	0.179 ± 0.076	-	-	0.107 ± 0.048	-	-
VoxelMorph	0.319 ± 0.151	0.386 ± 0.195	0.362 ± 0.165	0.233 ± 0.044	0.724 ± 0.627	0.524 ± 0.368
Res-U-Net	0.409 ± 0.282	0.452 ± 0.369	0.393 ± 0.190	0.881 ± 0.772	1.007 ± 0.913	0.854 ± 0.756
LK-U-Net	0.440 ± 0.407	0.382 ± 0.177	0.388 ± 0.227	0.859 ± 0.737	0.852 ± 0.696	0.891 ± 0.780
TransMorph	0.450 ± 0.354	0.464 ± 0.322	0.443 ± 0.298	0.930 ± 0.817	0.375 ± 0.211	0.308 ± 0.152
Dual-PR-Net	0.346 ± 0.102	0.372 ± 0.138	0.353 ± 0.097	0.583 ± 0.529	0.224 ± 0.039	0.288 ± 0.137
LapIRN	0.614 ± 0.444	-	-	1.096 ± 1.018	-	-
Median magnitude						
Method	Inter-subject test set (N=100)			Intra-subject test set (N=7)		
	a	b	c	a	b	c
Elastix	16.00 ± 5.12	-	-	15.20 ± 3.44	-	-
VoxelMorph	11.29 ± 2.90	12.11 ± 3.21	13.55 ± 3.87	10.02 ± 2.41	11.37 ± 2.59	13.26 ± 2.70
Res-U-Net	13.10 ± 3.37	13.96 ± 3.83	15.11 ± 4.17	11.89 ± 2.93	12.93 ± 2.90	14.65 ± 2.99
LK-U-Net	11.22 ± 3.20	13.37 ± 3.72	14.28 ± 4.10	9.65 ± 2.52	12.59 ± 2.66	13.93 ± 3.02
TransMorph	12.45 ± 3.03	14.02 ± 3.77	15.19 ± 4.22	11.85 ± 2.57	13.41 ± 2.69	14.97 ± 3.12
Dual-PR-Net	15.41 ± 4.16	15.89 ± 4.26	16.12 ± 4.51	14.74 ± 2.98	15.30 ± 3.28	15.72 ± 3.00
LapIRN	14.70 ± 4.08	-	-	14.15 ± 2.74	-	-

neck registration, thereby avoiding local minima driven merely by intensity matching and resulting in more plausible deformations. It is important to note that the difference between the two-stage MUSA approach and the two-stage multiresolution baseline is not readily apparent from the deformed images, as intensity matching is generally achieved to a reasonable extent.

5.2.3. Plausibility of deformations

The deformation regularity metrics, including folding percentage and $SDlog|J_\phi|$ are reported in Table 5. By comparing MUSA with the two-stage multiresolution baseline, we observed that the introduction of the MUSA loss did not result in any considerable increases in foldings or $SDlog|J_\phi|$. Therefore, the deformation regularity was maintained.

We also report the median value of the deformation magnitude in Table 5. Fig. 8 further shows the histogram of the deformation magnitude aggregated from all testing pairs. For architectures without explicit multiresolution modeling (i.e., VoxelMorph, Res-U-Net, LK-U-Net, and TransMorph), the single-stage registration resulted in reduced median magnitude and magnitude distribution shifting towards zero. This indicates that the intrinsic multiresolution modeling of U-Net and Swin Transformer was inadequate for capturing the large head-and-neck deformation. Both two-stage approaches improved large motion estimation. The two-stage MUSA approach further improved the large deformation recovery, especially for models with lower capacity. It also achieved more consistent deformation magnitude distribution across different architectures, as indicated by the reduced shaded areas in Fig. 8.

In addition to the differences in plausibility revealed by the deformation quiver plot in Fig. 7, we present the Jacobian determinant maps for the same cases in Fig. 9. The improvement in plausibility was more pronounced in the intra-subject case (Fig. 9b). Minimal volume change in bony structures is expected for intra-subject registration. Therefore, the Jacobian determinants should be close to one. For architectures without explicit multiresolution modeling, implausible values deviating from one were observed in the skull and mandible, in both the single-stage approach and the two-stage multiresolution approach. Dual-PR-Net and LapIRN showed improved plausibility due to their explicit multiresolution modeling, though LapIRN performed slightly worse than Dual-PR-Net. The two-stage MUSA approach consistently improved the deformation plausibility for models lacking explicit multiresolution modeling. The Jacobian determinants were close to one across the bony mask. However, there were still slight deviations for models with lower capacity. The effect of the MUSA approach also extended outside the bony structures, ensuring a more plausible Jacobian determinant map for the entire head-and-neck region. The arrows in the lower panel of Fig. 9b highlight the hot and cold spots representing unrealistic expansion and shrinkage, which has been mitigated by the MUSA approach. For the inter-subject case, the Jacobian determinant map was further complicated by the volume discrepancies between two subjects. Specifically, in the presented case, the moving subject had a thicker skull.

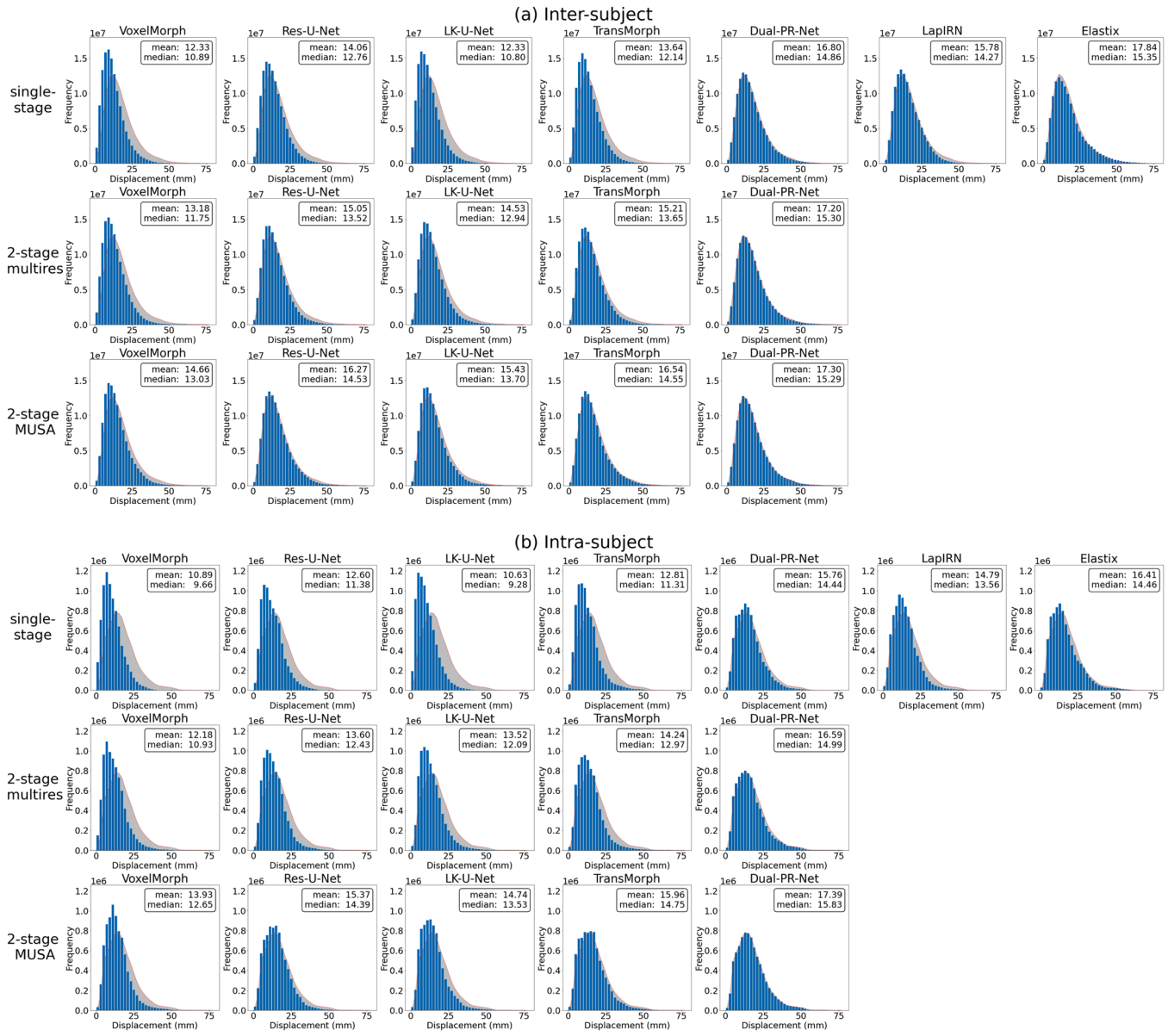

Fig. 8. Histograms of the deformation magnitude for (8a) inter-subject and (8b) intra-subject registration. These histograms were generated from all deformation vector fields in the test set, considering only deformation within the body mask. The two-stage MUSA approach using Dual-PR-Net architecture is selected as the reference, and its histogram profile is displayed in the red dotted line with gray shadings to highlight the difference. The mean and median values of the entire deformation magnitude distribution are shown in the top right corner of each subplot. The three rows correspond to single-stage registration, two-stage multi-resolution approach, and two-stage MUSA approach, respectively. The last three models (i.e., Dual-PR-Net, LapIRN and Elastix) has explicit multi-resolution modeling.

Jacobian determinant values greater than one should be expected within the skull mask. The MUSA approach yielded a more uniform and plausible skull expansion, whereas the two baseline methods showed over-estimated expansion concentrated near the frontal upper area of the head. For cases where the inter-patient bone size difference was relatively small, we observed similar Jacobian determinant patterns as in the intra-subject case discussed before.

5.2.4. *Effective receptive field analysis*

Figs. 10 and 11 present the ERF comparison. The ERF maps were rendered using maximum intensity projection (MIP) after taking the absolute value. Fig. 10 shows the ERFs of the two first-stage networks employed in the two-stage multiresolution baseline and two-stage MUSA approach for three representative architectures: VoxelMorph, TransMorph, and Dual-PR-Net. The probing point was positioned at the center of the image. A notable expansion of ERFs for the MUSA loss compared to the standard loss function was observed for both

VoxelMorph and TransMorph. This observation suggests that the proposed MUSA loss can enlarge the ERF of the network, which is vital not only for posture correction but also for encouraging a more plausible deformation by considering more anatomical context. The Dual-PR-Net exhibited a relatively large ERF even with the standard loss due to its explicit multiresolution design, explaining its superior performance.

Fig. 11 shows different ERFs with different values of α in MUSA loss. The probing point was set within the skull mask at the top of the head in the fixed image. As α increases, the ERF expands and encompasses more regions corresponding to the skull. The ERF extends to cover the most part of the skull when $\alpha = 1000$. This not only justifies the choice of α in the MUSA loss, but also demonstrates that the MUSA loss successfully introduces valid anatomical context to the registration task.

5.3. Comparison with Dice loss

The TransMorph architecture was selected to demonstrate the difference between the proposed MUSA loss and Dice loss as it was one of

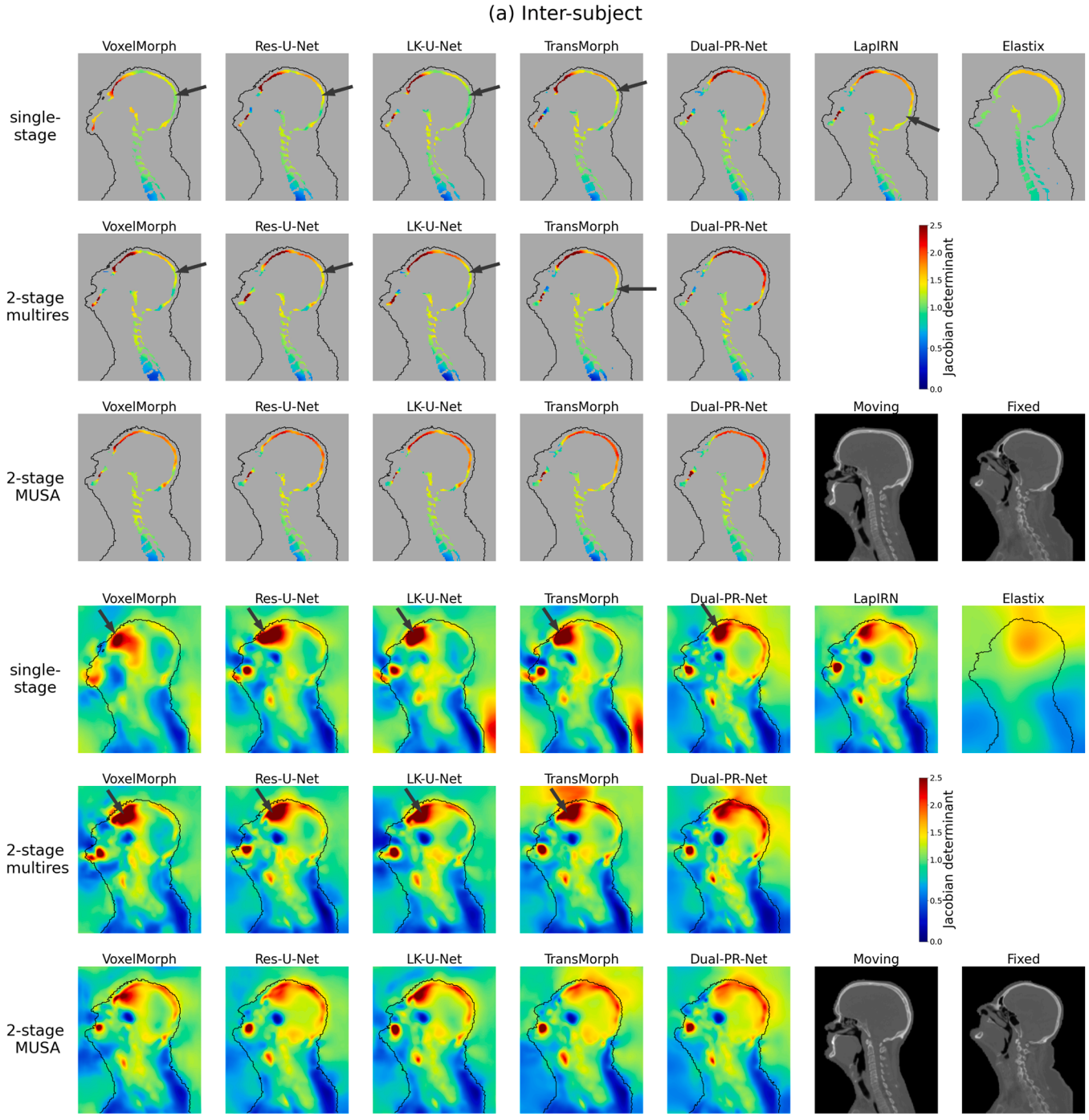

Fig. 9. Jacobian determinant maps of the deformation for (9a) inter-subjection and (9b) intra-subject registration. The same cases as in Fig. 7 are presented. The central slice of the coronal view is displayed. In the top panel, the maps are masked with the bony masks of the deformed image. In the bottom panel, the full Jacobian determinant maps are shown with the body contour of the deformed image. Arrows highlight implausible volume changes. The last three models (i.e., Dual-PR-Net, LapIRN and Elastix) has explicit multiresolution modeling.

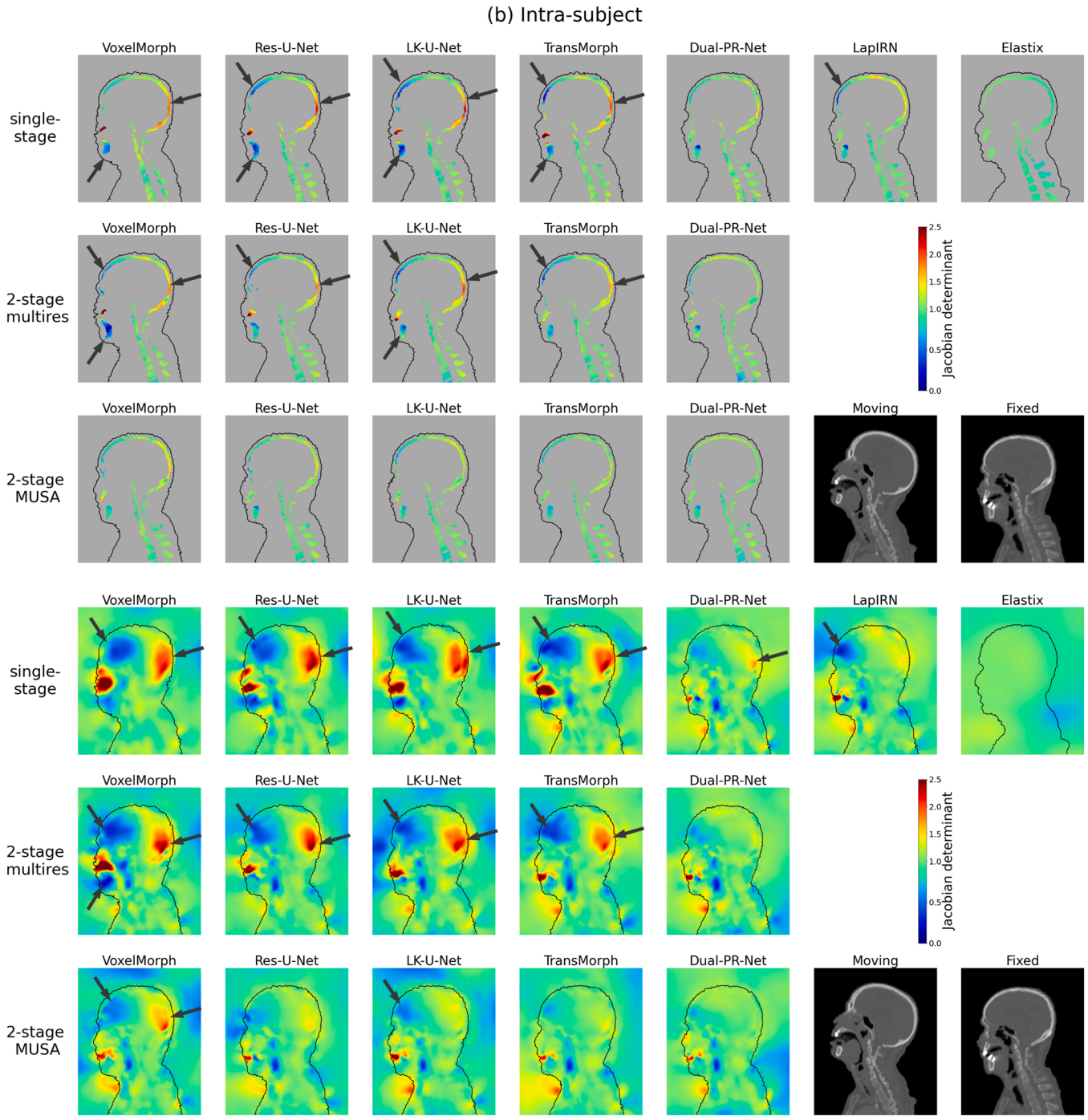

Fig. 9. (*continued*).

the more capable models but still showed distinct improvements from the MUSA framework, unlike the Dual-PR-Net. The comparison of the two-stage multiresolution baseline, the two-stage Dice loss approach, and the proposed two-stage MUSA approach is detailed in Appendix B.2. Table B1 summarizes the quantitative metrics including TRE and Dice. Fig. B2 shows the qualitative results as well as the Jacobian determinant maps. Incorporating Dice loss resulted in a slight increase in TRE. The observed bony Dice improved notably, while the Dice for unobserved structures remained unaffected. On the other hand, the MUSA loss reduced the TRE but only marginally affected the Dice scores for both observed and unobserved structures. The deformed images showed no distinct difference. However, the deformation quiver plot and the Jacobian determinant maps demonstrated that adding the Dice loss did not resolve the implausibility issue of the two-stage multiresolution baseline. These findings indicate that the MUSA loss and the Dice loss assist the registration in quite different ways.

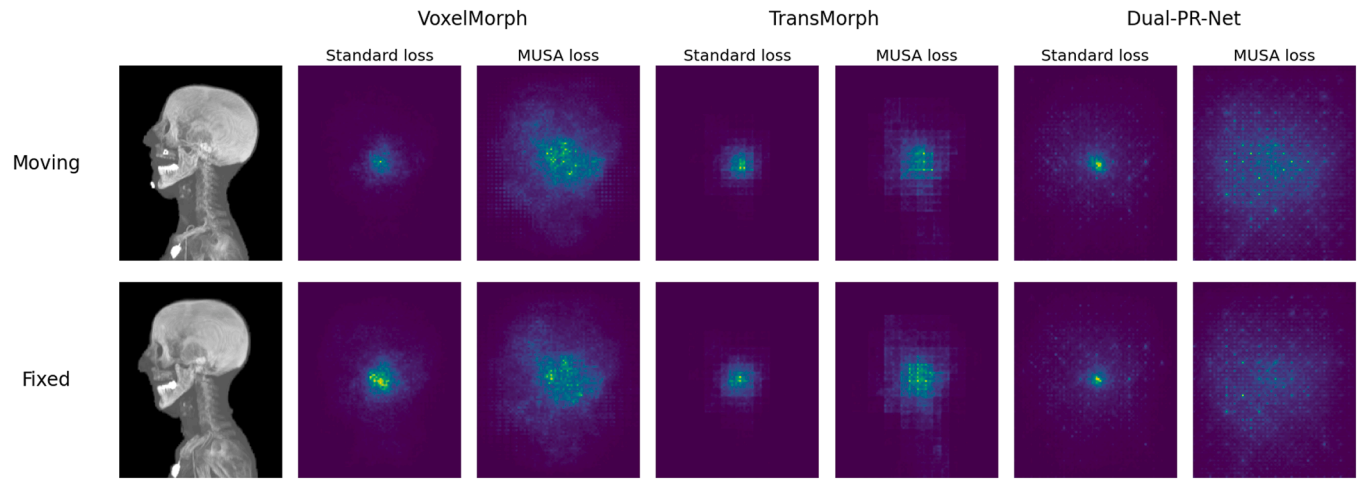

Fig. 10. The effect of the proposed MUSA loss on the effective receptive fields (ERFs). VoxelMorph, TransMorph, and Dual-PR-Net are shown as examples. For each architecture, the ERF is calculated for the first-stage networks used in the two-stage multiresolution approach (standard loss) and two-stage MUSA approach (MUSA loss). The probing point for ERF calculation is set at the center of the deformation field. The two rows show backpropagation to the moving and the fixed images, respectively. Maximum intensity projection (MIP) rendering along the coronal view is used for ERF visualization.

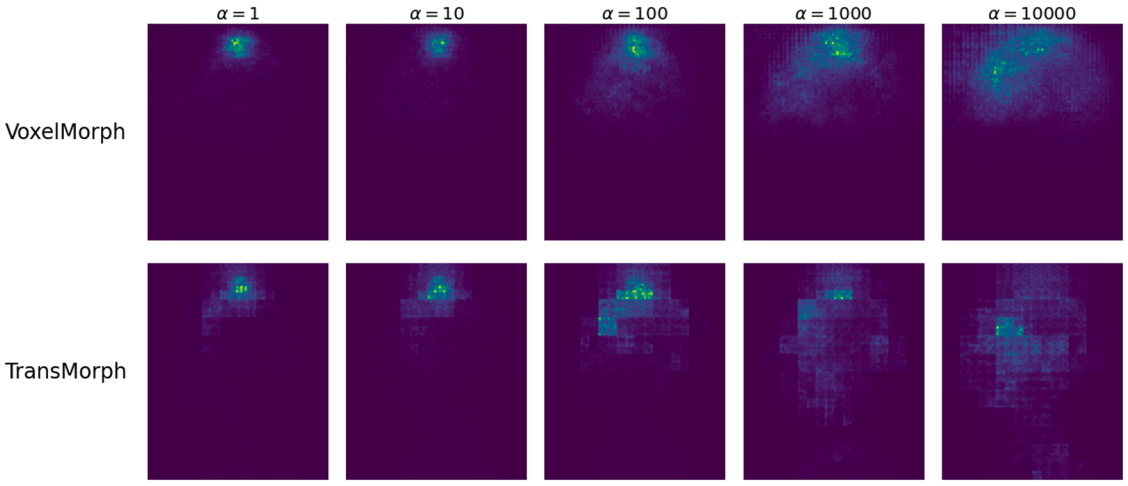

Fig. 11. ERFs corresponding to different α values in the MUSA loss (Eq. (5)) are demonstrated using VoxelMorph and TransMorph as examples. The probing point is set within the skull mask in the fixed image located at the top of the head. The ERFs corresponding to the moving image are shown. The result is similar for the fixed image. Maximum intensity projection (MIP) rendering along the coronal view is used for ERF visualization.

6. Discussion

This paper presents a MUsculo-Skeleton-Aware (MUSA) framework to anatomically guide head-and-neck CT registration. We have demonstrated consistent improvement across various architectures in registration accuracy with respect to Target Registration Error (TRE) and deformation plausibility.

Accurately and thoroughly evaluating deformable registration remains challenging due to the unknown ground truth (Viergever et al., 2016). We first discuss the implications and limitations of our evaluation results in this paper. It is well known that good intensity matching does not guarantee accurate or plausible deformations, as multiple voxels of similar intensities can be mapped to a target voxel to achieve a reasonable intensity match (Rohlfing, 2012; Wang et al., 2022). Alternatively, contour matching metrics have been widely used as surrogates and provide clinically relevant information in applications like radiation therapy. Although these metrics account for some anatomical information, they do not directly assess the underlying deformation and can be ambiguous within the contour boundary. Ideally, TRE provides an accurate metric for registration evaluation. However, generating a large number of landmarks in the head-and-neck region is labor-intensive and also challenging due to anatomical ambiguity, particularly for inter-subject registration. As a compromise, we used a relatively small set of 11 landmarks in our study. We deliberately selected landmarks that were distinct and easily recognizable to minimize annotation variability. The MUSA framework consistently achieved TRE

improvements across different architectures. The effect size decreased with increased model capacity, which was expected as more capable models would show diminishing returns. In addition, statistical testing on the large inter-subject test set showed statistically significant improvements with very low p-values, indicating that the improvement in TRE is real and not confounded by other factors such as labeling variability. Nevertheless, labeling variability remains a limitation of the current study. This also explains why the effect size of TRE improvement is larger for the intra-subject dataset than for the inter-subject dataset, as the former has smaller labeling variability due to closer anatomy and less complex deformation. Improving and automating the labeling process and curating a head-and-neck registration dataset with more densely labeled landmarks will be invaluable for future studies. This could potentially further highlight the contribution of the MUSA framework. Currently, the relatively small effect size in TRE improvement suggests that its clinical impact requires further validation.

To highlight the contribution of the proposed MUSA framework, we employed several ways to demonstrate the improved deformation plausibility. In addition to the commonly used deformation regularity metrics such as folding percentage and $SDlog|J_\phi|$, we analyzed deformation quiver plots, deformation magnitude distributions, and Jacobian determinant maps. These results further distinguished our proposed method from the two-stage multiresolution baseline. The MUSA approach better recovered large deformations, faithfully captured posture changes like head pitch, and alleviated unrealistic expansion and shrinkage for all architectures without explicit multiresolution modeling (i.e., VoxelMorph, Res-U-Net, LK-U-Net, and TransMorph). Architectures with explicit multiresolution modeling (i.e., Dual-PR-Net and LapIRN) showed better plausibility. LapIRN still showed some unrealistic deformation in the single-stage setting. Dual-PR-Net predicted plausible deformations in all three settings and benefited the least from incorporating the MUSA framework. This is probably because Dual-PR-Net has a large ERF from its multiresolution design and predicts multi-level deformation on feature pyramids, which contain richer information than image pyramids. Nevertheless, it is safe to conclude that the proposed MUSA framework can be confidently integrated with applicable architectures to improve registration plausibility. The plausibility evaluation methods employed in this study are still limited as they only reveal implausibility related to large and global deformations. Developing other evaluation metrics that can reflect implausibility on a finer scale or of other types is of interest for future studies.

We attribute MUSA's improvement to a divide-and-conquer strategy that predicts the complex head-and-neck deformation by sequentially addressing the bulk posture change and residual fine-scale deformation. This provides a simplified yet reasonable approximation of decomposing the head-and-neck deformation into musculoskeletal motion and residual tissue deformation, without pursuing detailed biomechanical modeling. In the MUSA framework, the Pos-Net adopts spatially variant regularization to account for the rigidity difference between the bony structures and soft tissue. With more stringent regularization, the deformed image only achieved partial matching with the fixed image, while the bone shapes were reasonably preserved. This resulted in an overall effect of posture alignment. The posture corrected image is much closer to the fixed image, and residual deformation can be more robustly addressed by the Ref-Net. In contrast, directly predicting the large and complex deformation may cause the registration algorithm to get trapped in a local minimum, especially for models with lower capacity. This results in the implausible deformation in the single-stage and two-stage baselines. By employing the divide-and-conquer strategy, MUSA enables models of varying capacities to achieve more consistent and improved registration results.

In addition, the performance gain of the MUSA framework can be partially explained by the enlarged ERF. Simply increasing the ERF through architectural design without considering anatomical context may be inadequate, as evidenced by the suboptimal performance of LK-U-Net and TransMorph in baseline settings. In MUSA loss, the increased bending energy on bony structures introduces strong coupling of the predicted deformations at different spatial locations within the same bone mask. This occurs because bending energy penalizes the second-order spatial gradient of the deformation. The bones in the head-and-neck region are large structures, particularly the skull and the mandible. We believe the network is enforced to expand its ERF in order to successfully predict such long-range dependencies in the deformation.

The MUSA framework also offers a more transparent and interpretable understanding of the registration process. Specifically, the two-stage decomposition provides an opportunity to identify and rectify the problematic stage in case of unsatisfactory registration. In general, the first stage tends to be more robust and reliable as it focuses on the high-contrast bony structures in CT images and is also more regularized. Residual soft tissue deformation can introduce more uncertainties due to the pronounced elasticity of soft tissue and insufficient contrast in CT images. Therefore, even if the second stage refinement encounters challenges, the posture correction can still provide valuable information and serve as a foundation for further refinement.

Although both MUSA loss and Dice loss utilize segmentation information, they do so in fundamentally different ways, as reflected in different behaviors on the evaluation metrics. Dice loss directly enforces segmentation matching, which improves observed Dice scores but does not enhance deformation plausibility or reduce TRE. In contrast, MUSA loss explicitly regularizes the deformation within bony masks, leading to improved deformation plausibility but only marginal gains in contour matching metrics. This indicates that the improvement in TRE from MUSA loss results from enhanced deformation plausibility rather than segmentation matching due to potential information leakage. Unfortunately, the improved plausibility does not translate to distinct improvement in segmentation matching. This is not unexpected as segmentation overlap may only weakly correlate with registration accuracy and plausibility, especially for large, unlocalized structures (Rohlfing, 2012). Further combining the MUSA framework with Dice loss could be preferable for certain tasks, but it is beyond the scope of the current study.

Our experiments also highlight the significance of employing an explicit multiresolution strategy in deep-learning-based DIR for handling large deformations. This is evident from the failure of VoxelMorph, Res-U-Net, LK-U-Net, and TransMorph in the single-stage setting and their notable improvement when adopting a two-level image pyramid with either the two-stage multiresolution or the two-stage MUSA approach. The deformation magnitude distribution in Fig. 8 indicates that this failure is related to underestimating large displacements. In contrast, methods with explicit multiresolution settings, including Dual-PR-Net, LapIRN, and Elastix, recovered the large deformation more faithfully and resulted in more plausible registration results. In the MUSA framework, the explicit two-level multiresolution strategy is seamlessly integrated into the two-stage decomposition, as the bulk posture change primarily corresponds to global motions, while the residual tissue deformation occurs at a finer resolution. Previous studies demonstrated that large motion in brain registration could be better addressed with models with larger ERFs through the use of either

larger kernels in CNNs (Jia et al., 2022) or the Transformer architecture (Chen et al., 2022). However, when applied to the more challenging task of head-and-neck registration, these methods remain inadequate. This indicates that the large deformation in the head-and-neck region exceeds the capacity of the implicit multiresolution modeling of U-Net and Swin Transformer (Hering et al., 2021). Therefore, the explicit multi-resolution strategy, including an image or feature pyramid and deformation refinement from coarse to fine, might be more important than architectural improvements and should be consistently employed for head-and-neck registration. This is also crucial for ensuring a fair comparison between different network architectures in deep-learning-based DIR.

Our method can accommodate both inter-subject and intra-subject head-and-neck CT registration despite their different characteristics and clinical applications. Intra-subject registration is in general a better-defined and less complex task with straightforward applications. For example, it is frequently used in radiation therapy for motion management, automatic re-contouring, radiation dose accumulation, and longitudinal treatment response evaluation. Inter-subject registration is more challenging due to significant anatomical differences among individuals and larger deformation ranges. These anatomical differences further complicate registration evaluation and clinical interpretation. However, inter-subject registration opens the door to broader, potentially more impactful studies. Currently, it is most often used in atlas-based segmentation (Sims et al., 2009). Emerging applications in population outcome analysis, such as normal tissue complication probability (NTCP) modeling (Monti et al., 2018; Palma et al., 2019) could be of great clinical value. Due to the scarcity of paired intra-subject datasets, we trained our networks exclusively on the inter-patient dataset. For testing, we used a larger inter-subject dataset and a relatively small intra-patient dataset. We intentionally curated an intra-patient dataset with large deformation from different patient setup protocols (Hwang et al., 2009), which is substantially more challenging compared to studies using the same patient setup (Lei et al., 2022). Although the intra-subject test set is small, it provides a more controlled testing scenario to highlight the contribution of MUSA (e.g., Fig 9b). Our results also showed a greater improvement in TRE for intra-subject compared to inter-subject test sets. This is partly due to increased landmark labeling uncertainty in inter-subject data. Additionally, intra-subject registration involves less tissue deformation, with musculoskeletal motion contributing more to the total deformation. Since the Pos-Net in MUSA framework is specifically designed to address musculoskeletal motion, the more distinct improvement on intra-subject dataset is expected. Nevertheless, for a pure intra-subject registration task, enforcing stricter rigidity constraints remains preferable to ensure anatomical accuracy. Such work necessitates a large, curated intra-subject dataset for training and is beyond the scope of the current study.

Our work has several limitations and could benefit from improvements in future research. The MUSA framework is limited by its simplified anatomical prior, focusing solely on bony structures and soft tissue. Integrating additional biomechanical knowledge from muscle and joint modeling in the head-and-neck region (Alizadeh et al., 2020; Lavallee et al., 2013) can be helpful to further improve registration accuracy and plausibility. Further improvement in bone segmentation can help increase the performance of the proposed MUSA framework. Differentiating between upper and lower teeth was challenging for some patients in our study, occasionally leading to the lower teeth merging with the skull mask (e.g., the first two patients in Fig. 2) and resulting in unrealistic deformations. Moreover, metal artifacts caused by dental fillings, which are prevalent in head-and-neck CT scans, can severely degrade registration performance. The proposed method and all the methods tested in this study face significant challenges in the presence of metal artifacts. Addressing these artifacts is of great clinical importance and will require specific strategies in future work, such as suppressing artifacts before registration or incorporating metal mask information within the registration framework. Another unsolved problem in head-and-neck CT registration is related to topological changes caused by mouth opening and closing. None of the presented registration methods can correctly create or eliminate the air gap due to mouth opening and closing when there is topological difference between the fixed and moving images. Handling such an issue likely requires more detailed segmentation of the oral cavity substructures, and specialized techniques to deal with content mismatch and topological changes (e.g., Alderliesten et al. (2013); Nithiananthan et al. (2012)), which is beyond the scope of the current study.

7. Conclusion

This study presents a MUsculo-Skeleton-Aware (MUSA) framework for deep-learning-based deformable image registration, targeting the challenging task of head-and-neck CT registration. The proposed Pos-Net + Ref-Net framework decomposes the complex head-and-neck deformation into a bulk posture change and residual fine deformation and then tackles them sequentially. The efficacy of the proposed framework has been demonstrated across various network architectures, consistently improving both registration accuracy and deformation plausibility. Our experiments highlight the significance of explicit multiresolution modeling and anatomical guidance in head-and-neck CT registration to ensure anatomically plausible deformations.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT (OpenAI) in order to check grammar and improve readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Hengjie Liu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Elizabeth McKenzie:** Data curation, Software, Visualization, Writing – review & editing. **Di Xu:** Methodology, Validation, Writing – review & editing. **Qifan Xu:** Methodology, Software, Validation. **Robert K. Chin:** Data curation, Validation, Writing – review & editing, Formal analysis. **Dan Ruan:** Conceptualization, Formal analysis, Methodology, Project administration, Supervision, Validation, Writing – review & editing. **Ke Sheng:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ke Sheng reports financial support was provided by National Institutes of Health. Ke Sheng reports financial support was provided by the U.S. Office of Congressionally Directed Medical Research Programs.

Data availability

The inter-subject datasets can be obtained through The Cancer Imaging Archive (TCIA), but we cannot redistribute them. We do not have the permission to share the intra-subject dataset.

Acknowledgements

We thank The Cancer Imaging Archive (TCIA) for providing the datasets. This work was supported by grants from the National Institutes of Health (U.S.) under grant numbers R44CA183390, R01EB031577, and 75N91019C00053, as well as the U.S. Department of Defense under grant W81XWH2210044.

Appendix A. Hyperparameter tuning

A.1. Tuning of the smoothness regularization weight λ

We performed independent hyperparameter tuning for all the single-stage and two-stage baselines used in our study to determine the optimal λ in Eq. (3), considering both Dice and TRE on the validation set. The models were trained for fewer epochs for hyperparameter tuning than the final models to conserve computational resources. However, convergence was confirmed. The single-stage models underwent training for 500 epochs. The two-stage framework's low-resolution component was also trained for 500 epochs. Subsequently, the weights for the low-resolution network were frozen. We then integrated the second-stage network, initialized it with weights from the single-stage training, and performed additional training for 200 epochs. The best Dice and TRE were recorded and reported.

The tuning curves are shown in Fig. A1. Both the Dice and TRE curves exhibited relatively flat regions near the optimal value; however, the optimal values of λ for Dice and TRE could differ by a factor of 2 to 5. In regions where the Dice plateaus, our selection criterion prioritized the TRE metric. This is because TRE is a stronger indicator of registration accuracy, as segmentation overlap described by Dice score is an indirect surrogate of the underlying deformation, which is less effective in distinguishing reasonable from poor registration (Rohlfing, 2012). For Dual-PR-Net and LapIRN, both curves were relatively flat for $\lambda \leq 0.1$, in which case we chose the optimal λ of 0.1, since it resulted in better deformation regularity without harming the accuracy. The optimal λ selected for each architecture and framework is summarized in Table A1.

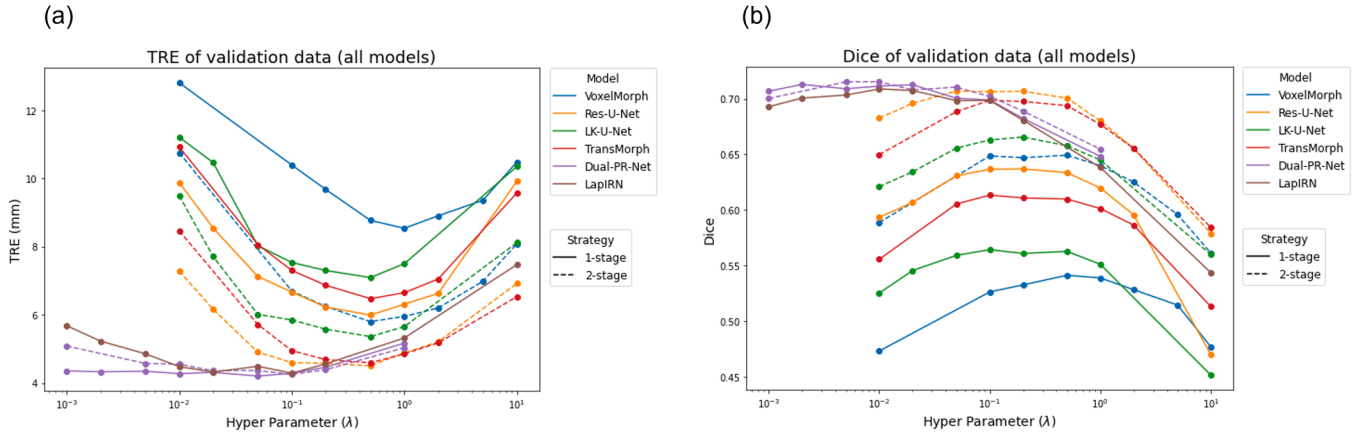

Fig. A1. (a) TRE and (b) Dice scores of validation data for all single-stage (solid lines) and two-stage (dotted lines) baselines used in the experiment with varying λ values in Eq. (3).

Table A1
The optimal smoothness hyperparameter (λ) used for the baseline models in the experiments.

Method	Single-stage	Two-stage
VoxelMorph	1.0	0.5
Res-U-Net	0.5	0.5
LK-U-Net	0.5	0.5
TransMorph	0.5	0.5
Dual-PR-Net	0.1	0.1
LapIRN	0.1	-

A.2. Tuning of α in MUSA loss

For MUSA loss (i.e., Eq. (5)), we fixed λ to be the optimal value from the two-stage multiresolution baseline and tuned α independently, as explained in Section 4.3. The tuning for α was conducted on three representative architectures with varying capacity, including VoxelMorph, TransMorph, and Dual-PR-Net. We followed the same training strategy as the two-stage multiresolution baseline. Note that when $\alpha = 1$ in MUSA framework, it is the same as the two-stage multiresolution baseline.

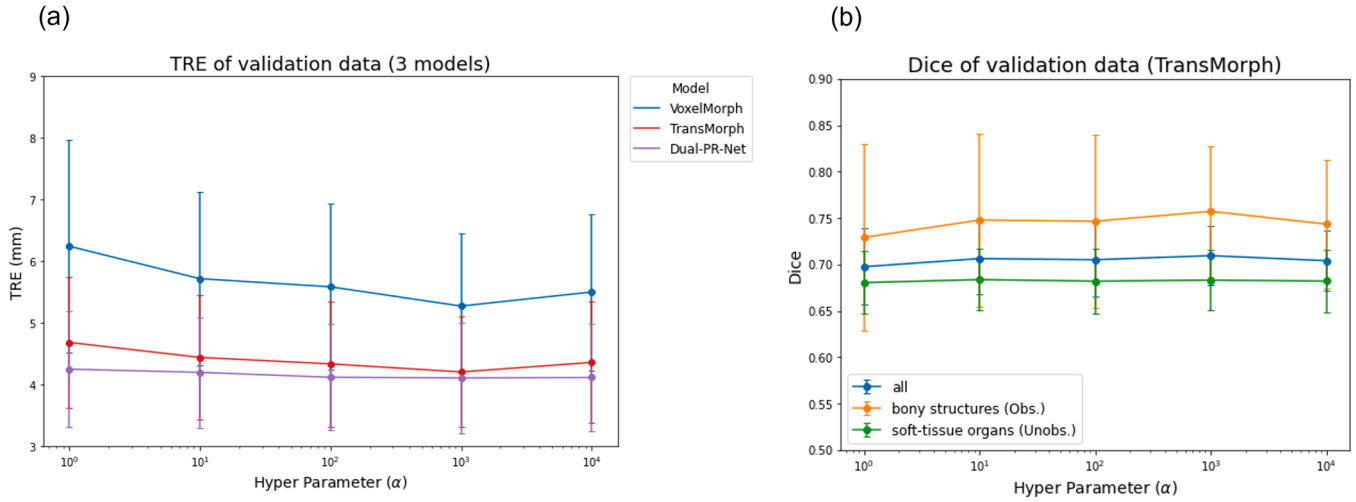

Fig. A2. (a) TRE of validation data given different α values in MUSA loss (Eq. (5)) for three representative models: VoxelMorph, TransMorph, and Dual-PR-Net. (b) Dice scores of validation data for TransMorph given different α values in MUSA loss. The three curves show Dice scores for all structures, observed bony structures, and unobserved soft tissue organs, respectively. Note that in both figures, the leftmost point of each curve (i.e., $\alpha = 1$) corresponds to the two-stage multiresolution baseline (i.e., baseline (b)) trained with the standard loss function (Eq. (3)).

Fig. A2(a) shows the TRE outcomes on the validation dataset. All three models showed best TRE results when $\alpha = 1000$. The consistent behavior of α across different models with varying registration capacities is expected, as α is attempting to characterize the relative rigidity of bony structures compared to soft tissue and thus should be relatively invariant. However, we observed diminishing gains with the increase of model capacity. VoxelMorph, with the smallest capacity indicating by its worst performance at $\alpha = 1$, showed the largest improvement at the optimal α value. TransMorph demonstrated moderate improvement with the integration of MUSA loss. Dual-PR-Net showed negligible improvement. The Dice results for TransMorph are shown in Fig. A2(b). Since we only used the bony segmentations in MUSA loss, we refer to bony segmentations as observed and the soft tissue organs as unobserved and plot them in addition to the average Dice score of all structures. The MUSA loss only slightly improved the Dice of observed bony structures, while the Dice for unobserved soft tissue organs were not affected by α value. As a result, the optimal α value of 1000 was selected based on the TRE metric alone.

Appendix B. Comparison between MUSA and Dice loss

B.1. Tuning of γ in Dice loss

For Dice loss (i.e., Eq. (6)), we similarly fixed λ to be the optimal value from the two-stage baseline and tuned γ independently. The tuning results of γ with respect to TRE and Dice are shown in Fig. B1(a,c). An optimal γ of 10^{-4} was selected for Dice loss, as it improved the Dice score of observed bony structures without significantly compromising the unobserved Dice of soft tissue organs or TRE.

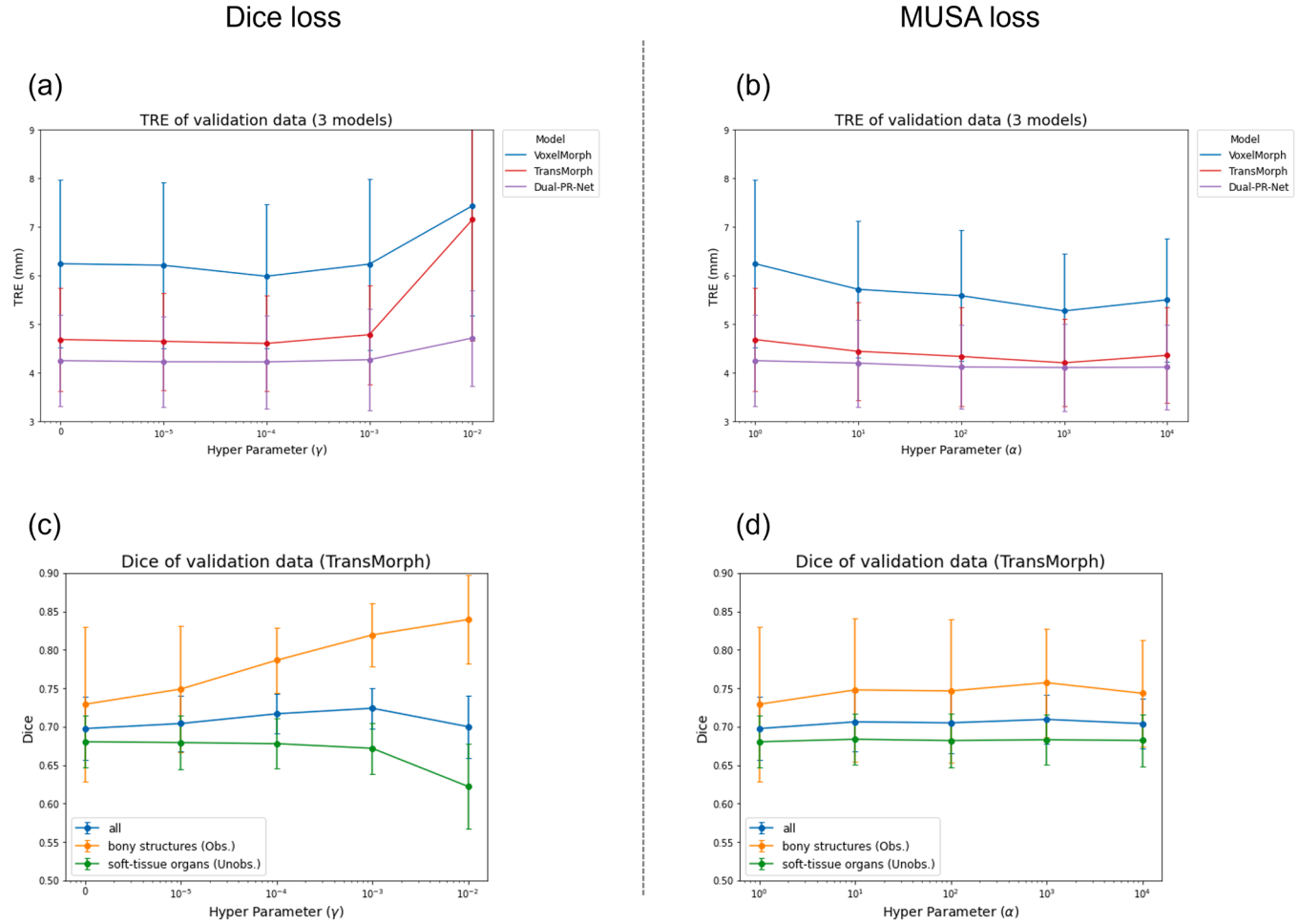

Fig. B1. TRE and Dice scores of validation data given different γ values in Dice loss and different α values in MUSA loss. (a) TRE for different γ values in Dice loss, (b) TRE for different α values in MUSA loss, (c) Dice scores for different γ values in Dice loss, (d) Dice scores for different α values in MUSA loss. Note that the leftmost point in each curve (i.e., $\gamma = 0$ in Dice loss or $\alpha = 1$ in MUSA loss) corresponds to the two-stage baseline trained with the standard loss function (Eq. (3)).

We also show a side-by-side comparison of the tuning curves for Dice loss and MUSA loss to facilitate the understanding of MUSA's contribution. As the hyperparameter γ increased, there was a notable improvement in the Dice score for observed bony structures; however, the Dice score for unobserved soft tissue organs declined when γ exceeds 10^{-4}. This observation was consistent with the previous study by Balakrishnan et al. (2019). Conversely, varying α only showed a marginal enhancement in the Dice score for bony structures up to an optimal point, beyond which the score deteriorated, while the Dice score for unobserved structures remained stable. As for TRE, the introduction of Dice loss did not lead to any improvement, and excessive γ values might be detrimental. On the other hand, increasing α resulted in improvement in TRE up to an optimal value. The tuning curves revealed the fundamental differences between the Dice loss and the MUSA loss, despite that they both incorporated the bony segmentations in the loss function.

B.2. Results of Dice loss

To demonstrate the difference between the proposed MUSA loss and Dice loss, we present the Dice loss results by a comparison with the two-stage multiresolution baseline and the two-stage MUSA approach. The TransMorph architecture was used for all the experiments. Table B1 shows the quantitative metrics, including TRE and Dice scores. The Dice scores were categorized into the average Dice of all structures, observed bony Dice and unobserved soft tissue organ Dice to better understand the impact of each loss functions. Fig. B2 shows the qualitative results including the deformed images and the deformation quiver plots. We also show the Jacobian determinant maps for plausibility analysis. The same inter- and intra-subject cases were used as in Figs. 7 and 9. The results and their implications were described in Section 5.3.

Table B1
To demonstrate the difference between MUSA loss and Dice loss, we present a quantitative comparison between the two-stage multiresolution baseline, the two-stage Dice loss baseline, and the two-stage MUSA approach. The average target registration error (TRE in mm) and average Dice scores (including Dice for all structures, observed bony structures and unobserved soft tissue structures) are reported. The best result within each column is highlighted in **bold**. The TransMorph architecture was used for all experiments.

Inter-subject test set (N=100)				
Framework	TRE (mm) ↓	Dice (all) ↑	Dice (bone) ↑	Dice (soft) ↑
2-stage multiresolution	5.04 ± 1.76	0.698 ± 0.041	0.735 ± 0.086	**0.678 ± 0.037**
2-stage multiresolution + Dice	5.12 ± 1.78	**0.718 ± 0.027**	**0.797 ± 0.035**	0.677 ± 0.036
2-stage MUSA	**4.61 ± 1.32**	0.703 ± 0.037	0.750 ± 0.068	0.674 ± 0.037
Intra-subject test set (N=7)				
Framework	TRE (mm) ↓	Dice (all) ↑	Dice (bone) ↑	Dice (soft) ↑
2-stage multiresolution	3.26 ± 0.85	0.754 ± 0.032	0.847 ± 0.022	0.702 ± 0.041
2-stage multiresolution + Dice	3.41 ± 0.79	**0.759 ± 0.028**	**0.867 ± 0.011**	0.699 ± 0.039
2-stage MUSA	**2.64 ± 0.34**	0.755 ± 0.029	0.843 ± 0.022	**0.705 ± 0.038**

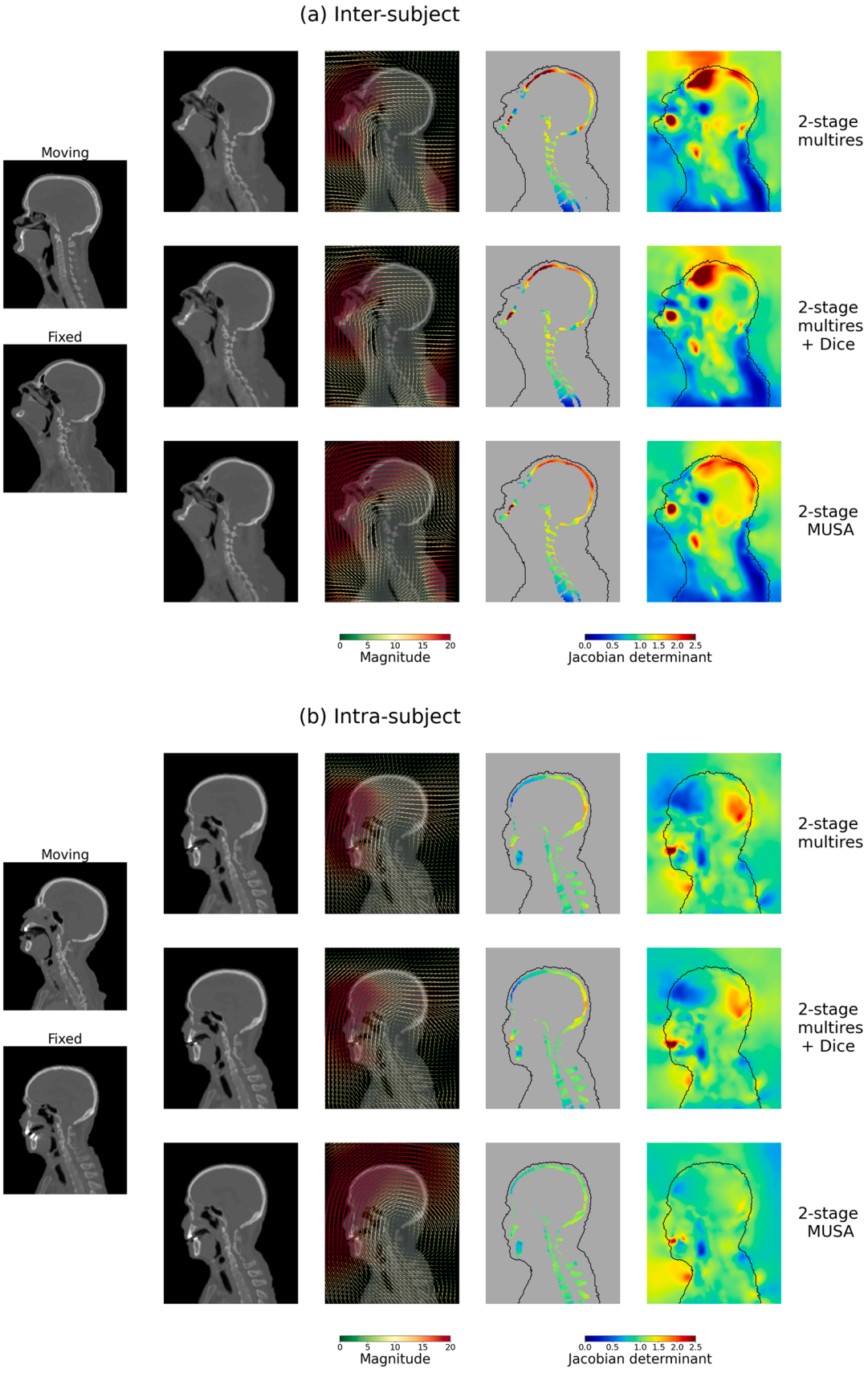

Fig. B2. To demonstrate the difference between MUSA loss and Dice loss, we present a comparison between the two-stage multiresolution baseline, the two-stage Dice loss baseline, and the two-stage MUSA approach. The same cases as in Figs. 7 and 9 are shown. The qualitative registration results, similar to those in Fig. 7, include deformed images and deformation quiver plots. Jacobian determinant maps, similar to those in Fig. 9, are also shown. The TransMorph architecture was used for all experiments.

References

Aerts, H.J.W.L., Velazquez, E.R., Leijenaar, R.T.H., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M.M., Leemans, C.R., Dekker, A., Quackenbush, J., Gillies, R.J., Lambin, P., 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. 5, 4006. https://doi.org/10.1038/ncomms5006.

Alderliesten, T., Sonke, J.-J., Bosman, P.A.N., 2013. Deformable image registration by multi-objective optimization using a dual-dynamic transformation model to account for large anatomical differences. Medical Imaging 2013: Image Processing. Presented at the Medical Imaging 2013: Image Processing, SPIE, pp. 273–279. https://doi.org/10.1117/12.2006783.

Alizadeh, M., Knapik, G.G., Mageswaran, P., Mendel, E., Bourekas, E., Marras, W.S., 2020. Biomechanical musculoskeletal models of the cervical spine: A systematic literature review. Clin. Biomech. 71, 115–124. https://doi.org/10.1016/j.clinbiomech.2019.10.027.

Al-Mayah, A., Moseley, J., Hunter, S., Velec, M., Chau, L., Breen, S., Brock, K., 2010. Biomechanical-based image registration for head and neck radiation treatment. Phys. Med. Biol. 55, 6491–6500. https://doi.org/10.1088/0031-9155/55/21/010.

Ang, K.K., Zhang, Q., Rosenthal, D.I., Nguyen-Tan, P.F., Sherman, E.J., Weber, R.S., Galvin, J.M., Bonner, J.A., Harris, J., El-Naggar, A.K., Gillison, M.L., Jordan, R.C., Konski, A.A., Thorstad, W.L., Trotti, A., Beitler, J.J., Garden, A.S., Spanos, W.J., Yom, S.S., Axelrod, R.S., 2014. Randomized Phase III Trial of Concurrent Accelerated Radiation Plus Cisplatin With or Without Cetuximab for Stage III to IV Head and Neck Carcinoma: RTOG 0522. J. Clin. Oncol. https://doi.org/10.1200/JCO.2013.53.5633.

Armstrong, R.A., 2014. When to use the Bonferroni correction. Ophthalmic Physiol. Opt. 34, 502–508. https://doi.org/10.1111/opo.12131.

Arsigny, V., Commowick, O., Pennec, X., Ayache, N., 2006. A Log-Euclidean Framework for Statistics on Diffeomorphisms. In: Larsen, R., Nielsen, M., Sporring, J. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 924–931. https://doi.org/10.1007/11866565_113.

Bajcsy, R., Kovačič, S., 1989. Multiresolution Elastic Matching. Comput. Vis. Graphic. Image Process. 46 (1), 1–21.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Trans. Med. Imaging 38, 1788–1800. https://doi.org/10.1109/TMI.2019.2897538.

Beichel, R.R., Ulrich, E.J., Bauer, C., Wahle, A., Brown, B., Chang, T., Plichta, K., Smith, B., Sunderland, J., Braun, T., Fedorov, A., Clunie, D., Onken, M., Magnotta, V. A., Menda, Y., Riesmeier, J., Pieper, S., Kikinis, R., Graham, M.M., Casavant, T.L., Sonka, M., Buatti, J., 2015. Data From QIN-HEADNECK (Version 4) [Data set]. Cancer Imaging Arch.

Bejarano, T., De Ornelas-Couto, M., Mihaylov, I., 2018. Head-and-neck squamous cell carcinoma patients with CT taken during pre-treatment, mid-treatment, and post-treatment (HNSCC-3DCT-RT) [Data set]. Cancer Imaging Arch.

Bejarano, T., De Ornelas-Couto, M., Mihaylov, I.B., 2019. Longitudinal fan-beam computed tomography dataset for head-and-neck squamous cell carcinoma patients. Med. Phys. 46, 2526–2537. https://doi.org/10.1002/mp.13460.

Bharatha, A., Hirose, M., Hata, N., Warfield, S.K., Ferrant, M., Zou, K.H., Suarez-Santana, E., Ruiz-Alzola, J., D'Amico, A., Cormack, R.A., Kikinis, R., Jolesz, F.A., Tempany, C.M.C., 2001. Evaluation of three-dimensional finite element-based deformable registration of pre- and intraoperative prostate imaging. Med. Phys. 28, 2551–2560. https://doi.org/10.1118/1.1414009.

Bosch, W.R., Straube, W.L., Matthews, J.W., Purdy, J.A., 2015. Head-Neck Cetuximab [Data set]. Cancer Imaging Arch.

Boveiri, H.R., Khayami, R., Javidan, R., Mehdizadeh, A., 2020. Medical image registration using deep neural networks: A comprehensive review. Comput. Electr. Eng. 87, 106767. https://doi.org/10.1016/j.compeleceng.2020.106767.

Brock, K.K., Sharpe, M.B., Dawson, L.A., Kim, S.M., Jaffray, D.A., 2005. Accuracy of finite element model-based multi-organ deformable image registration: Accuracy of FEM-based multi-organ deformable image registration. Med. Phys. 32, 1647–1659. https://doi.org/10.1118/1.1915012.

Cao, X., Yang, J., Zhang, J., Wang, Q., Yap, P.-T., Shen, D., 2018. Deformable Image Registration Using a Cue-Aware Deep Regression Network. IEEE Trans. Biomed. Eng. 65, 1900–1911. https://doi.org/10.1109/TBME.2018.2822826.

Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y., 2021. ViT-V-Net: Vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468.

Chen, J., Liu, Y., He, Y., Du, Y., 2023a. Spatially-varying Regularization with Conditional Transformer for Unsupervised Image Registration. arXiv preprint arXiv:2303.06168.

Chen, J., Liu, Y., Wei, S., Bian, Z., Subramanian, S., Carass, A., Prince, J.L., Du, Y., 2023b. A survey on deep learning in medical image registration: new technologies, uncertainty, Evaluation Metrics, and Beyond. https://doi.org/10.48550/arXiv.2307.15615.

Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y., 2022. TransMorph: Transformer for unsupervised medical image registration. Med. Image Anal. 82, 102615. https://doi.org/10.1016/j.media.2022.102615.

Chi, Y., Liang, J., Yan, D., 2006. A material sensitivity study on the accuracy of deformable organ registration using linear biomechanical modelsa). Med. Phys. 33, 421–433. https://doi.org/10.1118/1.2163838.

Christensen, G.E., Johnson, H.J., 2001. Consistent image registration. IEEE Trans. Med. Imaging 20, 568–582. https://doi.org/10.1109/42.932742.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F., 2013. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J. Digit. Imaging 26, 1045–1057. https://doi.org/10.1007/s10278-013-9622-7.

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2019. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Med. Image Anal. 57, 226–236. https://doi.org/10.1016/j.media.2019.07.006.

Dauguet, J., Herard, A.-S., Declerck, J., Delzescaux, T., 2009. Locally constrained cubic B-spline deformations to control volume variations. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. Presented at the 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 983–986. https://doi.org/10.1109/ISBI.2009.5193219.

de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., 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, pp. 204–212.

Ding, X., Zhang, X., Zhou, Y., Han, J., Ding, G., Sun, J., 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs. arXiv preprint arXiv:2203.06717.

Ding, Z., Niethammer, M., 2022. Aladdin: Joint Atlas Building and Diffeomorphic Registration Learning with Pairwise Alignment. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, New Orleans, LA, USA, pp. 20752–20761. https://doi.org/10.1109/CVPR52688.2022.02012.

du Bois d'Aische, A., De Craene, M., Macq, B., Warfield, S.K., 2005a. An Improved Articulated Registration Method For Neck Images. In: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. Presented at the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. IEEE, Shanghai, China, pp. 7668–7671. https://doi.org/10.1109/IEMBS.2005.1616288.

du Bois d'Aische, A., de Craene, M., Macq, B., Warfield, S.K., 2005b. An articulated registration method. In: IEEE International Conference on Image Processing 2005. Presented at the 2005 International Conference on Image Processing. IEEE, Genova, Italy, pp. I–21. https://doi.org/10.1109/ICIP.2005.1529677.

Elhalawani, H., Mohamed, A.S.R., White, A.L., Zafereo, J., Wong, A.J., Berends, J.E., AboHashem, S., Williams, B., Aymard, J.M., Kanwar, A., Perni, S., Rock, C.D., Cooksey, L., Campbell, S., Ding, Y., Lai, S.Y., Marai, E.G., Vock, D., Canahuate, G.M., Freymann, J., Farahani, K., Kalpathy-Cramer, J., Fuller, C.D., MICCAI/M.D. Anderson Cancer Center Head and Neck Quantitative Imaging Working Group, 2017. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. Sci. Data 4, 170077. https://doi.org/10.1038/sdata.2017.77.

Eppenhof, K.A..J., Lafarge, M.W., Moeskops, P., Veta, M., Pluim, J.P.W., 2018. Deformable image registration using convolutional neural networks. In: Angelini, E. D., Landman, B.A. (Eds.), Medical Imaging 2018: Image Processing. Presented at the Image Processing. SPIE, Houston, United States, p. 27. https://doi.org/10.1117/12.2292443.

Eppenhof, K.A.J., Lafarge, M.W., Veta, M., Pluim, J.P.W., 2020. Progressively Trained Convolutional Neural Networks for Deformable Image Registration. IEEE Trans. Med. Imaging 39, 1594–1604. https://doi.org/10.1109/TMI.2019.2953788.

Fedorov, A., Clunie, D., Ulrich, E., Bauer, C., Wahle, A., Brown, B., Onken, M., Riesmeier, J., Pieper, S., Kikinis, R., Buatti, J., Beichel, R.R., 2016. DICOM for quantitative imaging biomarker development: a standards based approach to sharing clinical data and structured PET/CT analysis results in head and neck cancer research. PeerJ 4, e2057. https://doi.org/10.7717/peerj.2057.

Fischer, B., Modersitzki, J., 2003. Curvature Based Image Registration. J. Math. Imaging Vis. 18, 81–85. https://doi.org/10.1023/A:1021897212261.

Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X., 2020. Deep learning in medical image registration: a review. Phys. Med. Biol. 65, 20TR01. https://doi.org/10.1088/1361-6560/ab843e.

Fu, Y., Lei, Y., Wang, T., Patel, P., Jani, A.B., Mao, H., Curran, W.J., Liu, T., Yang, X., 2021. Biomechanically constrained non-rigid MR-TRUS prostate registration using deep learning based 3D point cloud matching. Med. Image Anal. 67, 101845. https://doi.org/10.1016/j.media.2020.101845.

Gerig, T., Shahim, K., Reyes, M., Vetter, T., Lüthi, M., Golland, P., Hata, N., Barillot, C., Hornegger, J., 2014. Spatially Varying Registration Using Gaussian Processes. In: Howe, R. (Ed.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. Springer International Publishing, Cham, pp. 413–420. https://doi.org/10.1007/978-3-319-10470-6_52.

Greer, H., Kwitt, R., Vialard, F.-X., Niethammer, M., 2021. ICON: Learning Regular Maps Through Inverse Consistency. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Montreal, QC, Canada, pp. 3376–3385. https://doi.org/10.1109/ICCV48922.2021.00338.

Grossberg, A., Elhalawani, H., Mohamed, A., Mulder, S., Williams, B., White, A., Zafereo, J., Wong, A., Berends, J., AboHashem, S., Aymard, J., Kanwar, A., Perni, S., Rock, C., Chamchod, S., Kantor, M., Browne, T., Hutcheson, K., Gunn, G., Frank, S., Rosenthal, D., Garden, A., Fuller, C., 2020. M.D. Anderson Cancer Center Head and Neck Quantitative Imaging Working Group. (2020) HNSCC [Dataset]. Cancer Imaging Arch.

Grossberg, A.J., Mohamed, A.S.R., Elhalawani, H., Bennett, W.C., Smith, K.E., Nolan, T. S., Williams, B., Chamchod, S., Heukelom, J., Kantor, M.E., Browne, T., Hutcheson, K.A., Gunn, G.B., Garden, A.S., Morrison, W.H., Frank, S.J., Rosenthal, D. I., Freymann, J.B., Fuller, C.D., 2018. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. Sci. Data 5, 180173. https://doi.org/10.1038/sdata.2018.173.

Haskins, G., Kruger, U., Yan, P., 2020. Deep learning in medical image registration: a survey. Mach. Vis. Appl. 31, 8. https://doi.org/10.1007/s00138-020-01060-x.

He, Y., Anderson, B.M., Cazoulat, G., Rigaud, B., Almodovar-Abreu, L., Pollard-Larkin, J., Balter, P., Liao, Z., Mohan, R., Odisio, B., Svensson, S., Brock, K.K., 2023. Optimization of mesh generation for geometric accuracy, robustness, and efficiency of biomechanical-model-based deformable image registration. Med. Phys. 50, 323–329. https://doi.org/10.1002/mp.15939.

Heinrich, M.P., Hansen, L., 2022. VoxelMorph++ going beyond the cranial vault with keypoint supervision and multi-channel instance optimisation. In: Biomedical Image Registration: 10th International Workshop, WBIR 2022, Munich, Germany, July 10–12, 2022, Proceedings. Springer, pp. 85–95.

Hering, A., Häger, S., Moltz, J., Lessmann, N., Heldmann, S., van Ginneken, B., 2021. CNN-based lung CT registration with multiple anatomical constraints. Med. Image Anal. 72, 102139. https://doi.org/10.1016/j.media.2021.102139.

Hering, A., Hansen, L., Mok, T.C.W., Chung, A.C.S., Siebert, H., Hager, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., Vesal, S., Rusu, M., Sonn, G., Estienne, T., Vakalopoulou, M., Han, L., Huang, Y., Yap, P.-T., Brudfors, M., Balbastre, Y., Joutard, S., Modat, M., Lifshitz, G., Raviv, D., Lv, J., Li, Q., Jaouen, V., Visvikis, D., Fourcade, C., Rubeaux, M., Pan, W., Xu, Z., Jian, B., De Benetti, F., Wodzinski, M., Gunnarsson, N., Sjolund, J., Grzech, D., Qiu, H., Li, Z., Thorley, A., Duan, J., Grosbrohmer, C., Hoopes, A., Reinertsen, I., Xiao, Y., Landman, B., Huo, Y., Murphy, K., Lessmann, N., van Ginneken, B., Dalca, A.V., Heinrich, M.P., 2023. Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE Trans. Med. Imaging 42, 697–712. https://doi.org/10.1109/TMI.2022.3213983.

Hering, A., van Ginneken, B., Heldmann, S., 2019. mlVIRNET: Multilevel variational image registration network. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 257–265. https://doi.org/10.1007/978-3-030-32226-7_29.

Hipwell, J.H., Vavourakis, V., Han, L., Mertzanidou, T., Eiben, B., Hawkes, D.J., 2016. A review of biomechanically informed breast image registration. Phys. Med. Biol. 61, R1. https://doi.org/10.1088/0031-9155/61/2/R1.

Hu, B., Zhou, S., Xiong, Z., Wu, F., 2022. Recursive decomposition network for deformable image registration. IEEE J. Biomed. Health Inform. 26, 5130–5141. https://doi.org/10.1109/JBHI.2022.3189696.

Hu, X., Kang, M., Huang, W., Scott, M.R., Wiest, R., Reyes, M., 2019. Dual-stream pyramid registration network. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 382–390. https://doi.org/10.1007/978-3-030-32245-8_43.

Hu, Y., Gibson, E., Ghavami, N., Bonmati, E., Moore, C.M., Emberton, M., Vercauteren, T., Noble, J.A., Barratt, D.C., 2018a. Adversarial Deformation Regularization for Training Image Registration Neural Networks. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 774–782. https://doi.org/10.1007/978-3-030-00928-1_87.

Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C.M., Emberton, M., Ourselin, S., Noble, J.A., Barratt, D.C., Vercauteren, T., 2018b. Weakly-supervised convolutional neural networks for multimodal image registration. Med. Image Anal. 49, 1–13. https://doi.org/10.1016/j.media.2018.07.002.

Hwang, A.B., Bacharach, S.L., Yom, S.S., Weinberg, V.K., Quivey, J.M., Franc, B.L., Xia, P., 2009. Can Positron Emission Tomography (PET) or PET/Computed Tomography (CT) Acquired in a Nontreatment Position Be Accurately Registered to a Head-and-Neck Radiotherapy Planning CT? Int. J. Radiat. Oncol. 73, 578–584. https://doi.org/10.1016/j.ijrobp.2008.09.041.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods 18, 203–211. https://doi.org/10.1038/s41592-020-01008-z.

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025.

Jia, X., Bartlett, J., Zhang, T., Lu, W., Qiu, Z., Duan, J., 2022. U-Net vs Transformer: Is U-Net Outdated in Medical Image Registration? In: Lian, C., Cao, X., Rekik, I., Xu, X., Cui, Z. (Eds.), Machine Learning in Medical Imaging, Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp. 151–160. https://doi.org/10.1007/978-3-031-21014-3_16.

Jian, B., Azampour, M.F., De Benetti, F., Oberreuter, J., Bukas, C., Gersing, A.S., Foreman, S.C., Dietrich, A.-S., Rischewski, J., Kirschke, J.S., Navab, N., Wendler, T., 2022. Weakly-Supervised Biomechanically-Constrained CT/MRI Registration of the Spine. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp. 227–236. https://doi.org/10.1007/978-3-031-16446-0_22.

Kabus, S., Franz, A., Fischer, B., 2006. Variational Image Registration with Local Properties. In: Pluim, J.P.W., Likar, B., Gerritsen, F.A. (Eds.), Biomedical Image Registration. Springer, Berlin, Heidelberg, pp. 92–100. https://doi.org/10.1007/11784012_12.

Kang, M., Hu, X., Huang, W., Scott, M.R., Reyes, M., 2022. Dual-stream pyramid registration network. Med. Image Anal. 78, 102379. https://doi.org/10.1016/j.media.2022.102379.

Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.-G., Ye, J.C., 2021. CycleMorph: Cycle consistent unsupervised deformable image registration. Med. Image Anal. 71, 102036. https://doi.org/10.1016/j.media.2021.102036.

Kim, J., Matuszak, M.M., Saitou, K., Balter, J.M., 2013. Distance-preserving rigidity penalty on deformable image registration of multiple skeletal components in the neck: Distance-preserving rigidity penalty for deformable image registration. Med. Phys. 40, 121907. https://doi.org/10.1118/1.4828783.

Kim, J., Saitou, K., Matuszak, M.M., Balter, J.M., 2016. A finite element head and neck model as a supportive tool for deformable image registration. Int. J. Comput. Assist. Radiol. Surg. 11, 1311–1317. https://doi.org/10.1007/s11548-015-1335-6.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J., 2010. elastix: A Toolbox for Intensity-Based Medical Image Registration. IEEE Trans. Med. Imaging 29, 196–205. https://doi.org/10.1109/TMI.2009.2035616.

Lavallee, A.V., Ching, R.P., Nuckley, D.J., 2013. Developmental biomechanics of neck musculature. J. Biomech. 46, 527–534. https://doi.org/10.1016/j.jbiomech.2012.09.029.

Lei, Y., Fu, Y., Tian, Z., Wang, T., Dai, X., Roper, J., Yu, D.S., McDonald, M., Bradley, J. D., Liu, T., Zhou, J., Yang, X., 2022. Deformable CT image registration via a dual feasible neural network. Med. Phys. 49, 7545–7554. https://doi.org/10.1002/mp.15875.

Lester, H., Arridge, S.R., 1999. A survey of hierarchical non-linear medical image registration. Pattern Recognit 32, 129–149. https://doi.org/10.1016/S0031-3203(98)00095-8.

Li, Z., Tian, L., Mok, T.C.W., Bai, X., Wang, P., Ge, J., Zhou, J., Lu, L., Ye, X., Yan, K., Jin, D., 2023b. SAMConvex: fast discrete optimization for ct registration using self-supervised anatomical embedding and correlation pyramid. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 559–569. https://doi.org/10.1007/978-3-031-43999-5_53.

Li, X., Zhang, Y., Shi, Y., Wu, S., Xiao, Y., Gu, X., Zhen, X., Zhou, L., 2017. Comprehensive evaluation of ten deformable image registration algorithms for contour propagation between CT and cone-beam CT images in adaptive head & neck radiotherapy. PLOS ONE 12, e0175906. https://doi.org/10.1371/journal.pone.0175906.

Li, J., Chen, J., Tang, Y., Wang, C., Landman, B.A., Zhou, S.K., 2023a. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. Med. Image Anal. 85, 102762. https://doi.org/10.1016/j.media.2023.102762.

Liang, X., Morgan, H., Nguyen, D., Jiang, S., 2021. Deep Learning–Based CT-to-CBCT Deformable Image Registration for Autosegmentation in Head and Neck Adaptive Radiation Therapy. J. Artif. Intell. Med. Sci. 2, 62–75. https://doi.org/10.2991/jaims.d.210527.001.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Montreal, QC, Canada, pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.

Loeckx, D., Maes, F., Vandermeulen, D., Suetens, P., 2004. Nonrigid Image Registration Using Free-Form Deformations with a Local Rigidity Constraint. In: Barillot, C., Haynor, D.R., Hellier, P. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 639–646. https://doi.org/10.1007/978-3-540-30135-6_78.

Luo, W., Li, Y., Urtasun, R., Zemel, R., 2016. Understanding the effective receptive field in deep convolutional neural networks. Proceedings of the 30th International Conference on Neural Information Processing Systems.

Maintz, J.B.A., Viergever, M.A., 1998. A survey of medical image registration. Med. Image Anal. 2, 1–36. https://doi.org/10.1016/S1361-8415(01)80026-8.

McCulloch, M.M., Anderson, B.M., Cazoulat, G., Peterson, C.B., Mohamed, A.S.R., Volpe, S., Elhalawani, H., Bahig, H., Rigaud, B., King, J.B., Ford, A.C., Fuller, C.D., Brock, K.K., 2019. Biomechanical modeling of neck flexion for deformable alignment of the salivary glands in head and neck cancer images. Phys. Med. Biol. 64, 175018. https://doi.org/10.1088/1361-6560/ab2f13.

McKenzie, E.M., Santhanam, A., Ruan, D., O'Connor, D., Cao, M., Sheng, K., 2020. Multimodality image registration in the head-and-neck using a deep learning-derived synthetic CT as a bridge. Med. Phys. 47, 1094–1104. https://doi.org/10.1002/mp.13976.

Mok, T.C.W., Chung, A.C.S., Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., 2020. Large deformation diffeomorphic image registration with laplacian pyramid networks. In: Joskowicz, L. (Ed.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 211–221. https://doi.org/10.1007/978-3-030-59716-0_21.

Monti, S., Pacelli, R., Cella, L., Palma, G., 2018. Inter-patient image registration algorithms to disentangle regional dose bioeffects. Sci. Rep. 8, 4915. https://doi.org/10.1038/s41598-018-23327-0.

National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC), 2018. The clinical proteomic tumor analysis consortium head and neck squamous cell carcinoma collection (CPTAC-HNSCC) (Version 12) [Data set]. Cancer Imaging Arch.

Neylon, J., Qi, X., Sheng, K., Staton, R., Pukala, J., Manon, R., Low, D.A., Kupelian, P., Santhanam, A., 2014. A GPU based high-resolution multilevel biomechanical head and neck model for validating deformable image registration: High-resolution biomechanical head and neck deformable phantoms. Med. Phys. 42, 232–243. https://doi.org/10.1118/1.4903504.

Niethammer, M., Kwitt, R., Vialard, F.-X., 2019. Metric Learning for Image Registration. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR). IEEE, Long Beach, CA, USA, pp. 8455–8464. https://doi.org/10.1109/CVPR.2019.00866.
Nithiananthan, S., Schafer, S., Mirota, D.J., Stayman, J.W., Zbijewski, W., Reh, D.D., Gallia, G.L., Siewerdsen, J.H., 2012. Extra-dimensional Demons: A method for incorporating missing tissue in deformable image registration. Med. Phys. 39, 5718–5731. https://doi.org/10.1118/1.4747270.
Pace, D.F., Aylward, S.R., Niethammer, M., 2013. A locally adaptive regularization based on anisotropic diffusion for deformable image registration of sliding organs. IEEE Trans. Med. Imaging 32, 2114–2126. https://doi.org/10.1109/TMI.2013.2274777.
Palma, G., Monti, S., Conson, M., Pacelli, R., Cella, L., 2019. Normal tissue complication probability (NTCP) models for modern radiation therapy. Semin. Oncol. 46, 210–218. https://doi.org/10.1053/j.seminoncol.2019.07.006.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.D, Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.
Payer, C., Štern, D., Bischof, H., Urschler, M., 2020. Coarse to Fine Vertebrae Localization and Segmentation with SpatialConfiguration-Net and U-Net. In: Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. Presented at the 15th International Conference on Computer Vision Theory and Applications. SCITEPRESS - Science and Technology Publications, Valletta, Malta, pp. 124–133. https://doi.org/10.5220/0008975201240133.
Qin, C., Wang, S., Chen, C., Qiu, H., Bai, W., Rueckert, D., Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., 2020. Biomechanics-informed neural networks for myocardial motion tracking in MRI. In: Joskowicz, L. (Ed.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 296–306. https://doi.org/10.1007/978-3-030-59716-0_29.
Rajagopal, V., Lee, A., Chung, J.-H., Warren, R., Highnam, R.P., Nielsen, P.M.F., Nash, M. P., 2007. Towards tracking breast cancer across medical images using subject-specific biomechanical models. In: Ayache, N., Ourselin, S., Maeder, A. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 651–658. https://doi.org/10.1007/978-3-540-75757-3_79.
Rohlfing, T., 2012. Image Similarity and Tissue Overlaps as Surrogates for Image Registration Accuracy: Widely Used but Unreliable. IEEE Trans. Med. Imaging 31, 153–163. https://doi.org/10.1109/TMI.2011.2163944.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
Ruan, D., Fessler, J.A., Roberson, M., Balter, J., Kessler, M., 2006. Nonrigid registration using regularization that accomodates local tissue rigidity, in: Reinhardt, J.M., Pluim, J.P.W. (Eds.). Presented at the Medical Imaging, San Diego, CA, p. 614412. https://doi.org/10.1117/12.653870.
Rueckert, D., Aljabar, P., Heckemann, R.A., Hajnal, J.V., Hammers, A., 2006. Diffeomorphic Registration Using B-Splines. In: Larsen, R., Nielsen, M., Sporring, J. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 702–709. https://doi.org/10.1007/11866763_86.
Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging 18, 712–721. https://doi.org/10.1109/42.796284.
Ruhaak, J., Polzin, T., Heldmann, S., Simpson, I.J.A., Handels, H., Modersitzki, J., Heinrich, M.P., 2017. Estimation of large motion in lung ct by integrating regularized keypoint correspondences into dense deformable registration. IEEE Trans. Med. Imaging 36, 1746–1757. https://doi.org/10.1109/TMI.2017.2691259.
Schnabel, J.A., Rueckert, D., Quist, M., Blackall, J.M., Castellano-Smith, A.D., Hartkens, T., Penney, G.P., Hall, W.A., Liu, H., Truwit, C.L., Gerritsen, F.A., Hill, D.L. G., Hawkes, D.J., 2001. A Generic Framework for Non-rigid Registration Based on Non-uniform Multi-level Free-Form Deformations. In: Niessen, W.J., Viergever, M.A. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 573–581. https://doi.org/10.1007/3-540-45468-3_69.
Sekuboyina, A., Husseini, M.E., Bayat, A., Löffler, M., Liebl, H., Li, H., Tetteh, G., Kukačka, J., Payer, C., Štern, D., Urschler, M., Chen, M., Cheng, D., Lessmann, N., Hu, Y., Wang, T., Yang, D., Xu, D., Ambellan, F., Amiranashvili, T., Ehlke, M., Lamecker, H., Lehnert, S., Lirio, M., Olaguer, N.P.de, Ramm, H., Sahu, M., Tack, A., Zachow, S., Jiang, T., Ma, X., Angerman, C., Wang, X., Brown, K., Kirszenberg, A., Puybareau, É., Chen, D., Bai, Y., Rapazzo, B.H., Yeah, T., Zhang, A., Xu, S., Hou, F., He, Z., Zeng, C., Xiangshang, Z., Liming, X., Netherton, T.J., Mumme, R.P., Court, L. E., Huang, Z., He, C., Wang, L.-W., Ling, S.H., Huỳnh, L.D., Boutry, N., Jakubicek, R., Chmelik, J., Mulay, S., Sivaprakasam, M., Paetzold, J.C., Shit, S., Ezhov, I., Wiestler, B., Glocker, B., Valentinitsch, A., Rempfler, M., Menze, B.H., Kirschke, J.S., 2021. VerSe: A Vertebrae labelling and segmentation benchmark for multi-detector CT images. Med. Image Anal. 73, 102166. https://doi.org/10.1016/j.media.2021.102166.
Sentker, T., Madesta, F., Werner, R., 2018. GDL-FIRE$$^\text {4D}$$: Deep Learning-Based Fast 4D CT Image Registration. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 765–773. https://doi.org/10.1007/978-3-030-00928-1_86.
Sermesant, M., Forest, C., Pennec, X., Delingette, H., Ayache, N., 2003. Deformable biomechanical models: Application to 4D cardiac image analysis. Med. Image Anal., Med. Image Comput. Comput. Assist. Interv. 7, 475–488. https://doi.org/10.1016/S1361-8415(03)00068-9.
Shi, J., He, Y., Kong, Y., Coatrieux, J.-L., Shu, H., Yang, G., Li, S., 2022. XMorpher: Full Transformer for Deformable Medical Image Registration via Cross Attention. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp. 217–226. https://doi.org/10.1007/978-3-031-16446-0_21.
Sims, R., Isambert, A., Grégoire, V., Bidault, F., Fresco, L., Sage, J., Mills, J., Bourhis, J., Lefkopoulos, D., Commowick, O., Benkebil, M., Malandain, G., 2009. A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. Radiother. Oncol. 93, 474–478. https://doi.org/10.1016/j.radonc.2009.08.013.
Sokooti, H., de Vos, B., Berendsen, F., Lelieveldt, B.P.F., Išgum, I., Staring, M., 2017. Nonrigid image registration using multi-scale 3d convolutional neural networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2017, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 232–239. https://doi.org/10.1007/978-3-319-66182-7_27.
Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable Medical Image Registration: A Survey. IEEE Trans. Med. Imaging 32, 1153–1190. https://doi.org/10.1109/TMI.2013.2265603.
Staring, M., Klein, S., Pluim, J.P.W., 2007. A rigidity penalty term for nonrigid registration: A rigidity penalty term for nonrigid registration. Med. Phys. 34, 4098–4108. https://doi.org/10.1118/1.2776236.
Teske, H., Bartelheimer, K., Meis, J., Bendl, R., Stoiber, E.M., Giske, K., 2017. Construction of a biomechanical head and neck motion model as a guide to evaluation of deformable image registration. Phys. Med. Biol. 62, N271–N284. https://doi.org/10.1088/1361-6560/aa69b6.
Tong, N., Gou, S., Yang, S., Cao, M., Sheng, K., 2019. Shape constrained fully convolutional DenseNet with adversarial training for multiorgan segmentation on head and neck CT and low-field MR images. Med. Phys. 46, 2669–2682. https://doi.org/10.1002/mp.13553.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is All you Need. Advances in Neural Information Processing Systems. Curran Associates, Inc.
Vialard, F.-X., Risser, L., 2014. Spatially-Varying Metric Learning for Diffeomorphic Image Registration: A Variational Framework. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. Springer International Publishing, Cham, pp. 227–234. https://doi.org/10.1007/978-3-319-10404-1_29.
Viergever, M.A., Maintz, J.B.A., Klein, S., Murphy, K., Staring, M., Pluim, J.P.W., 2016. A survey of medical image registration – under review. Med. Image Anal. 33, 140–144. https://doi.org/10.1016/j.media.2016.06.030.
Wang, D., Pan, Y., Durumeric, O.C., Reinhardt, J.M., Hoffman, E.A., Schroeder, J.D., Christensen, G.E., 2022. PLOSL: Population learning followed by one shot learning pulmonary image registration using tissue volume preserving and vesselness constraints. Med. Image Anal. 79, 102434. https://doi.org/10.1016/j.media.2022.102434.
Wang, Y., Qiu, H., Qin, C., 2023. Conditional deformable image registration with spatially-variant and adaptive regularization. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). Presented at the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). IEEE, Cartagena, Colombia, pp. 1–5. https://doi.org/10.1109/ISBI53787.2023.10230464.
Wee, L., Dekker, A., 2019. Data from HEAD-NECK-RADIOMICS-HN1 [Data set]. Cancer Imaging Arch.
Werner, R., Ehrhardt, J., Schmidt, R., Handels, H., 2009. Patient-specific finite element modeling of respiratory lung motion using 4D CT image data. Med. Phys. 36, 1500–1511. https://doi.org/10.1118/1.3101820.
Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Fast predictive multimodal image registration. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). Presented at the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, Melbourne, Australia, pp. 858–862. https://doi.org/10.1109/ISBI.2017.7950652.
Zhang, J., 2018. Inverse-consistent deep networks for unsupervised deformable image registration. arXiv preprint arXiv:1809.03443.
Zhang, Y., 2021. An unsupervised 2D–3D deformable registration network (2D3D-RegNet) for cone-beam CT estimation. Phys. Med. Biol. 66, 074001. https://doi.org/10.1088/1361-6560/abe9f6.
Zhao, S., Lau, T., Luo, J., Chang, E.I.-C., Xu, Y., 2020. Unsupervised 3D End-to-end medical image registration with volume tweening network. IEEE J. Biomed. Health Inform. 24, 1394–1404. https://doi.org/10.1109/JBHI.2019.2951024.
Zuley, M.L., Jarosz, R., Kirk, S., Lee, Y., Colen, R., Garcia, K., Delbeke, D., Pham, M., Nagy, P., Sevinc, G., Goldsmith, M., Khan, S., Net, J.M., Lucchesi, F.R., Aredes, N.D., 2016. The cancer genome atlas head-neck squamous cell carcinoma collection (TCGA-HNSC) (Version 6) [Data set]. Cancer Imaging Arch.