Improving Representativity of Conditional Femur Atlases via Perturbed-Condition Training

Chenqi Yang[a,b], Florence Dell'Aniello Picard[a], Pranav Poudel[a], and Herve Lombaert[a]

[a]Polytechnique Montreal, Montreal, QC, Canada
[b]University of Alberta, Edmonton, AB, Canada

ABSTRACT

Conditional atlases aim to represent phenotype-dependent anatomy, but conventional template and registration measures do not directly assess condition representativity. Condition representativity refers to how well the specified phenotype is reflected in the generated atlas. To improve this property, we propose perturbed-condition training. For the same target image, the method compares templates generated from matched and perturbed conditions, and encourages lower dissimilarity for the matched condition. The proportion of comparisons in which the matched condition yields lower dissimilarity is termed the matched-condition win rate. We evaluate the method on 1,120 femoral computed tomography (CT) volumes from the New Mexico Decedent Image Database[1] using 5-fold subject-level cross-validation, conditioning on age and weight. Compared with a conventional conditional-template baseline, our model increases a template Dice score from 0.871 to 0.903 and a post-registration Dice score from 0.885 to 0.929, while matched-condition win rates improve by 16.5-26.9 percentage points across the two attributes, with no observed deformation folding. These results suggest that perturbed-condition training improves condition representativity while preserving template and registration performance.

1. INTRODUCTION

Conditional atlases aim to capture anatomical variation associated with subject attributes by generating condition-specific templates. The conventional method generates templates from subject attributes and aligns them to individual subjects through deformable registration.[2] However, a conditional template may remain close to the population template, and the registration may compensate for this template–subject condition mismatch. Therefore, template similarity and registration accuracy may not fully reflect how well the supplied condition is represented.

Recent work has improved conditional atlas construction in many directions. Adversarial training has improved template image quality and sharpness.[3] CAS-Net jointly generates age-specific fetal-brain atlases and tissue segmentations.[4] Diffeomorphic formulations have modeled condition-dependent anatomical variation,[5] while AtlasMorph jointly learns intensity and label templates with improved atlas centrality.[6] However, these approaches primarily focus on atlas construction and registration quality. They do not explicitly enforce whether the supplied condition is represented in the generated atlas.

This paper introduces perturbed-condition training to improve condition representativity. For each target subject, we call its recorded phenotype the matched condition and a valid alternative phenotype the perturbed condition. The method encourages lower target dissimilarity under the matched condition than under the perturbed condition. We quantify the condition representativity using the matched-condition win rate, which is the rate at which the matched condition yields a lower dissimilarity. Our experiments evaluate femoral atlases conditioned on age and weight on CT data from the New Mexico Decedent Image Database (NMDID).[1,7]

2. METHOD

We propose perturbed-condition training to construct conditional atlases that better represent specified phenotypes. To this end, we define condition consistency. It requires that for a given subject, the matched condition yields lower target dissimilarity than a valid perturbed condition in both the generated template and its registered output (Fig. 1).

Further author information: (Send correspondence to H.L.)
H.L.: E-mail: herve.lombaert@polymtl.ca, Telephone: +1 (514) 340-4711

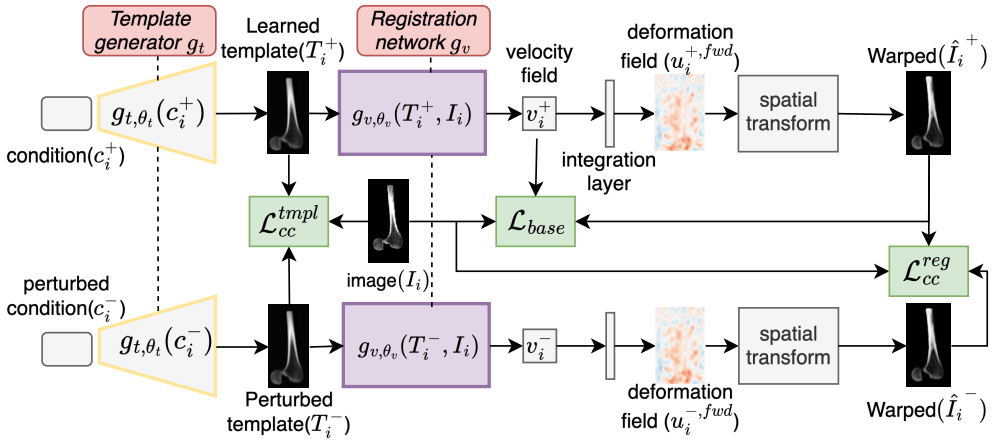


Figure 1. **Overview of perturbed-condition training.** Matched and perturbed conditions pass through shared template-generation and registration networks. The condition consistency objectives encourage the matched condition to yield lower dissimilarity to the target anatomy throughout template generation and registration.

2.1 Conditional Template Generation and Registration

Let $I_i \in \mathbb{R}^{H\times W\times D}$ denote the i-th femoral CT volume, and let $c_i \in \mathcal{C}$ denote its normalized condition vector, such as [age, weight]. Our conditional template follows a standard formulation,[2] defined as:

$$T_i \equiv T(c_i) = T_0 + g_t(c_i), \tag{1}$$

where T_0 is the fixed voxelwise mean and g_t predicts a condition-dependent residual. The shared decoder g_t produces the full-resolution residual from a trilinearly upsampled half-resolution feature map. The registration network[8,9] takes the conditional template and target image as input and predicts a stationary velocity field. Exponentiating this field gives the forward and inverse deformations:

$$v_i = g_v(T(c_i), I_i), \quad u_i^{\text{fwd}} = \exp(v_i), \quad u_i^{\text{inv}} = \exp(-v_i), \quad \widehat{I}_i = T(c_i) \circ u_i^{\text{fwd}}, \tag{2}$$

where u_i^{fwd} and u_i^{inv} are the forward and inverse deformations, and $\widehat{I}_i$ is the registered template. The base objective combines mean squared error (MSE) with centrality, smoothness, and deformation-magnitude terms:

$$\mathcal{L}_{\text{base}} = \text{MSE}(I_i, \widehat{I}_i) + \lambda_{\text{center}}\|\widetilde{u}\|_2^2 + \lambda_{\text{smooth}}\|\nabla u_i^{\text{fwd}}\|_2^2 + \lambda_{\text{magnitude}}\|u_i^{\text{fwd}}\|_2^2, \tag{3}$$

where the quantity $\widetilde{u}$ is a running estimate of the population mean inverse deformation. Its penalty follows unbiased atlas construction and keeps the template centered on the population.[2,10]

2.2 Perturbed-Condition Training

We improve condition representativity by maintaining condition consistency. Let c_i^+ be the recorded condition for subject i, called the **matched condition**. During training, a **perturbed condition** c_i^- is sampled uniformly within the valid attribute range, with a minimum normalized difference δ from the matched value for each changed continuous attribute. Both conditions are fed into the shared decoder g_t to produce templates $T_i^+ = T(c_i^+)$ and $T_i^- = T(c_i^-)$, which are independently registered to the same target image with the shared registration network:

$$\widehat{I}_i^+ = T_i^+ \circ \exp(g_v(T_i^+, I_i)), \qquad \widehat{I}_i^- = T_i^- \circ \exp(g_v(T_i^-, I_i)). \tag{4}$$

Two condition consistency objectives enforce lower image dissimilarity for T_i^+ than T_i^- and for $\widehat{I}_i^+$ than $\widehat{I}_i^-$:

$$\mathcal{L}_{\text{cc}}^{\text{tmpl}} = \max\big(0, m_{\text{tmpl}} + \text{MSE}(I_i, T_i^+) - \text{MSE}(I_i, T_i^-)\big), \tag{5}$$

$$\mathcal{L}_{\text{cc}}^{\text{reg}} = \max\Big(0, m_{\text{reg}} + \text{MSE}(I_i, \widehat{I}_i^+) - \text{MSE}(I_i, \widehat{I}_i^-)\Big), \tag{6}$$

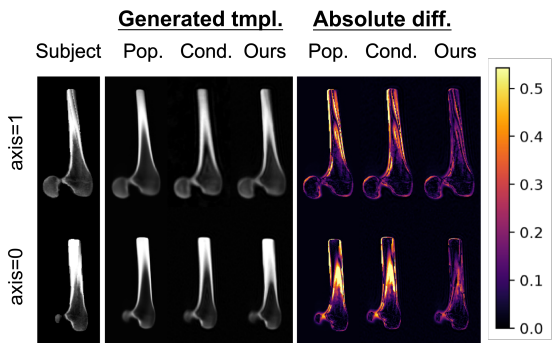


Figure 2. **Qualitative comparison for one subject.** Two orthogonal slices, templates from the population baseline, conditional baseline, and our method, and their absolute differences from the target. Our template shows smaller visual differences from the target than the baselines.

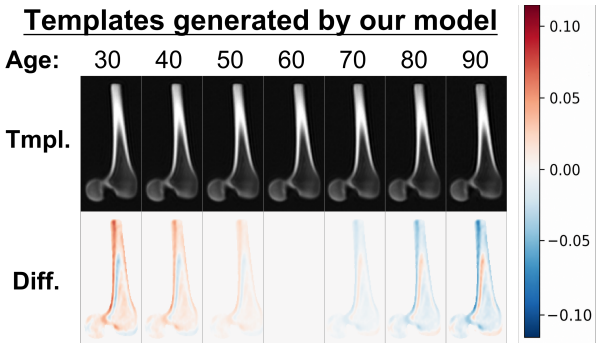


Figure 3. **Template variation with age.** Other attributes are fixed. Top: templates for ages 30 to 90 years. Bottom: differences from the 60-year-old template. Templates change progressively with age, indicating sensitivity to the conditioning attribute.

where m_{tmpl} and m_{reg} specify the minimum MSE advantage required for the matched condition at the template and registration stages, respectively. These condition consistency objectives encourage lower dissimilarities under the matched condition.

Registration can compensate for differences in the generated templates, allowing high registration accuracy to coexist with weak condition representativity. The two terms enforce condition consistency by working jointly on template generation and registration:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{tmpl}}\mathcal{L}_{\text{cc}}^{\text{tmpl}} + \lambda_{\text{reg}}\mathcal{L}_{\text{cc}}^{\text{reg}}. \tag{7}$$

3. EXPERIMENTS AND RESULTS

3.1 Experimental Setup

Our experiments evaluate whether perturbed-condition training improves condition representativity while preserving template similarity and registration performance. We used 1,120 foreground-masked femoral CT volumes from NMDID.[1,7] Five-fold subject-level cross-validation kept both femurs of each individual in the same fold. Each held-out fold contained 224 volumes, and Table 1 reports mean ± standard deviation across folds. Volumes were centered on the femoral foreground, cropped to $160 \times 192 \times 416$, and independently min-max normalized to $[0, 1]$. Age and weight were normalized using training-fold ranges.

We compared an unconditional population baseline, a conventional conditional-template baseline,[2] and our perturbed-condition training model. The conditional baseline and our model share the same template generator and registration network. Models were trained for 1,000 epochs using the Adam[11] optimizer with learning rate 10^{-4}, batch size 1, and 100 steps per epoch. For conditional models, the registration network was frozen for 400 epochs while the template generator used direct reconstruction, followed by 600 epochs of joint optimization. Checkpoints were selected by the lowest validation loss. The same preprocessing, folds and total number of training steps were used for all models. We set $\lambda_{\text{tmpl}} = \lambda_{\text{reg}} = 1.5$, $\lambda_{\text{center}} = 1$, $\lambda_{\text{smooth}} = 0.02$, $\lambda_{\text{magnitude}} = 0.01$, $m_{\text{tmpl}} = 3 \times 10^{-4}$, $m_{\text{reg}} = 4 \times 10^{-4}$ and $\delta = 0.2$.

Template similarity and registration performance were evaluated using the Dice coefficient (Dice), 95th-percentile Hausdorff distance (HD95), average symmetric surface distance (ASSD), and the percentage of voxels with a non-positive Jacobian determinant, reported as Fold. Condition representativity was evaluated with the **matched-condition win rate** and the mean matched-condition **error gaps**

$$\Delta_{\text{tmpl}} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \left[\text{MSE}(I_i, T_i^-) - \text{MSE}(I_i, T_i^+)\right], \quad \Delta_{\text{reg}} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \left[\text{MSE}(I_i, \widehat{I}_i^-) - \text{MSE}(I_i, \widehat{I}_i^+)\right]. \tag{8}$$

The win rate is the percentage of test pairs for which the matched condition gives lower error than the perturbed condition. A value above 50% indicates that the matched condition wins more often. A positive error gap

Table 1. **Template similarity and registration performance** for the population baseline, conditional baseline, and our method.

	Template Similarity			Registration Performance			
Method	Dice ↑	HD95 (mm) ↓	ASSD (mm) ↓	Dice ↑	HD95 (mm) ↓	ASSD (mm) ↓	Fold (%) ↓
Pop.	0.853±0.031	6.21±0.67	2.11±0.57	0.855±0.018	6.10±0.41	1.93±0.34	0.0000
Cond.	0.871±0.024	5.56±0.93	**1.73±0.28**	0.885±0.023	5.16±0.64	1.66±0.30	0.0000
Ours	**0.903±0.036**	**5.08±0.60**	1.79±0.17	**0.929±0.031**	**4.42±0.27**	**1.35±0.22**	0.0000

Table 2. **Condition representativity evaluation and loss ablation.** T-only and R-only retain only the template or registration contribution to the condition consistency objective, respectively.

(a) Weight: 40–120 kg

Metric	Cond.	Ours	T-only	R-only
Template win (%) ↑	60.1	**78.3**	63.5	55.1
Δ_{tmpl} ↑	8.54×10^{-4}	$\mathbf{1.26\times10^{-3}}$	5.92×10^{-4}	6.91×10^{-6}
Post-reg. win (%) ↑	65.9	**92.8**	60.1	91.9
Δ_{reg} ↑	3.93×10^{-4}	$\mathbf{1.65\times10^{-2}}$	1.03×10^{-6}	1.53×10^{-2}

(b) Age: 30–90 years

Metric	Cond.	Ours	T-only	R-only
Template win (%) ↑	65.1	**81.6**	64.7	57.9
Δ_{tmpl} ↑	1.38×10^{-3}	$\mathbf{2.50\times10^{-3}}$	4.50×10^{-5}	5.16×10^{-5}
Post-reg. win (%) ↑	62.5	**86.8**	52.0	83.6
Δ_{reg} ↑	3.12×10^{-4}	$\mathbf{1.14\times10^{-2}}$	3.26×10^{-7}	7.14×10^{-3}

indicates a mean advantage for the matched condition. Attribute-specific tests in Table 2 compare one evaluated attribute while holding the others fixed.

3.2 Template Similarity and Registration Performance

Our templates show smaller visual differences from the target than the population and conditional baseline templates (Fig. 2). When age varies and other attributes are fixed, the age sweep shows progressive template changes across the evaluated range (Fig. 3). This qualitative result complements the pairwise measures by showing that nearby conditions do not produce visually abrupt changes.

The conditional baseline improves most measures over the population baseline. Our model gives the highest template Dice, lowest template HD95, and best post-registration Dice, HD95, and ASSD (Table 1). The conditional baseline has a slightly lower template ASSD than our model. No folding was observed for any method.

3.3 Condition Representativity and Ablation

The conditional baseline improves average template and registration metrics over the population baseline, but matched-condition win rates are only 60.1–65.9% for the evaluated attributes. Together, Tables 1 and 2 indicate that average similarity and condition representativity capture different aspects of performance. For weight, our model raises template and post-registration win rates from 60.1% and 65.9% to 78.3% and 92.8%. For age, they rise from 65.1% and 62.5% to 81.6% and 86.8% (Table 2).

The post-registration-only variant (R-only) accounts for most of the post-registration gain, reaching 91.9% for weight and 83.6% for age. The template-only variant (T-only) does not reliably preserve its initial advantage, consistent with registration compensating for template differences. The combined model gives the highest win rates at both stages, suggesting complementary roles between $\mathcal{L}_{cc}^{tmpl}$ and $\mathcal{L}_{cc}^{reg}$.

4. CONCLUSION

We introduced perturbed-condition training to improve representativity in conditional femur atlases. Our method enforces condition consistency, requiring the matched condition to yield lower target dissimilarity throughout template generation and registration. We evaluated conditional femur atlases conditioned on age and weight using NMDID femoral CT. The proposed method improves most template and post-registration measures and increases matched-condition win rates, with no observed deformation folding. These results suggest that condition consistency improves representativity and that registration metrics alone do not fully capture this property. The current evaluation is limited to attribute values within the observed age and weight ranges. Future work will evaluate the method across additional anatomical structures and phenotype distributions, including cardiac and neuroimaging applications.

Disclosure: This manuscript is original and has not been previously published.

ACKNOWLEDGMENTS

The first author acknowledges scholarship support from the Polytechnique Montreal Research Internship Program and the support of PolyShape Lab. We acknowledge the New Mexico Decedent Image Database. The Free Access Decedent Database was funded by the National Institute of Justice under Grant No. 2016-DN-BX-0144.

REFERENCES

[1] Edgar, H. J. H., Daneshvari Berry, S., Moes, E., Adolphi, N. L., Bridges, P., and Nolte, K. B., "New Mexico Decedent Image Database," (2020).

[2] Dalca, A. V., Rakic, M., Guttag, J., and Sabuncu, M. R., "Learning conditional deformable templates with convolutional networks," in [*Advances in Neural Information Processing Systems*], **32**, 806–818, Curran Associates, Inc. (2019).

[3] Dey, N., Ren, M., Dalca, A. V., and Gerig, G., "Generative adversarial registration for improved conditional deformable templates," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*], 3929–3941 (October 2021).

[4] Li, L., Sinclair, M., Makropoulos, A., Hajnal, J. V., Edwards, A. D., Kainz, B., Rueckert, D., and Alansary, A., "CAS-Net: Conditional atlas generation and brain segmentation for fetal MRI," in [*Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*], *Lecture Notes in Computer Science* **12959**, 221–230, Springer (2021).

[5] Whitbread, L., Laurenz, S., Palmer, L. J., Jenkinson, M., and Alzheimer's Disease Neuroimaging Initiative, "Deep-diffeomorphic networks for conditional brain templates," *Human Brain Mapping* **46**(8), e70229 (2025).

[6] Rakic, M., Hoopes, A., Abulnaga, S. M., Sabuncu, M. R., Guttag, J. V., and Dalca, A. V., "AtlasMorph: Learning conditional deformable templates for brain MRI," *Medical Image Analysis* **110**, 103893 (2026).

[7] Daneshvari Berry, S. and Edgar, H. J. H., "Announcement: The New Mexico Decedent Image Database," *Forensic Imaging* **24**, 200436 (March 2021).

[8] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V., "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging* **38**(8), 1788–1800 (2019).

[9] Dalca, A. V., Balakrishnan, G., Guttag, J., and Sabuncu, M. R., "Unsupervised learning for fast probabilistic diffeomorphic registration," in [*Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*], *Lecture Notes in Computer Science* **11070**, 729–738, Springer International Publishing, Cham (2018).

[10] Joshi, S., Davis, B., Jomier, M., and Gerig, G., "Unbiased diffeomorphic atlas construction for computational anatomy," *NeuroImage* **23**(Supplement 1), S151–S160 (2004).

[11] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).