

CTCF: Cascaded Transformer with Cross-Attention and Super-Resolution for Unsupervised Medical Image Registration

Daniil V. Pasenko

Peter the Great St. Petersburg Polytechnic University
Saint Petersburg, Russia
daniil.pasenko@yandex.ru

Abstract—We present CTCF, a hybrid deep learning framework for unsupervised deformable medical image registration that combines transformer-based global correspondence modeling with super-resolution deformation reconstruction. The proposed method integrates a deformable cross-attention (DCA) transformer encoder with a multi-scale super-resolution decoder within a cascaded refinement architecture, enabling accurate alignment of anatomical structures while preserving deformation fidelity. To improve physical plausibility, CTCF incorporates inverse consistency, cycle consistency, and Jacobian-based regularization to suppress non-invertible transformations and encourage topology-preserving deformations. The method is evaluated on the OASIS 3D brain MRI dataset and compared against state-of-the-art transformer-based registration models, including TransMorph-DCA and UTSRMorph. Experimental results demonstrate that CTCF achieves competitive registration accuracy while substantially reducing folding artifacts and improving deformation regularity. These findings indicate that combining transformer-based encoders with super-resolution decoding and consistency-driven regularization provides a practical pathway toward more reliable and physically consistent medical image registration.

Keywords—*image registration; transformer; deformable cross-attention; super-resolution; inverse consistency; unsupervised learning.*

I. INTRODUCTION

Deformable medical image registration is a fundamental problem in medical image analysis, enabling spatial alignment of anatomical structures across subjects, imaging modalities, and time points. In brain MRI registration, accurate alignment requires modeling both global anatomical variability and fine-scale local deformations, while preserving topology and avoiding non-physical foldings of the deformation field [1], [2].

Recent learning-based approaches have substantially reduced the computational burden of deformable registration by directly predicting dense displacement fields from image pairs. Unsupervised deep learning frameworks, such as VoxelMorph, have demonstrated that competitive registration accuracy can be achieved without explicit optimization at inference time [3], [4]. Subsequent extensions have highlighted the importance of deformation regularity and stability, emphasizing that high overlap accuracy alone is insufficient if the resulting transformations violate anatomical topology [5].

More recently, transformer-based architectures have emerged as a powerful alternative to convolutional networks for medical image registration. By leveraging self-attention mechanisms, transformer-based models can capture long-

range spatial dependencies that are difficult to model using purely convolutional encoders [6]. TransMorph and its variants have shown that transformer-based encoders significantly improve global correspondence modeling in volumetric brain MRI [7]–[9]. In particular, deformable cross-attention mechanisms combined with hierarchical transformer backbones, such as Swin Transformers, enable sparse adaptive sampling of anatomically relevant regions while maintaining computational efficiency [10], [11].

Despite these advances, most transformer-based registration frameworks rely on conventional convolutional decoders with interpolation-based upsampling to recover full-resolution deformation fields. Such decoding strategies may limit the spatial fidelity of high-resolution displacement fields, especially when training is performed at reduced resolution. In parallel, physical consistency constraints—such as inverse consistency, cycle consistency, and Jacobian-based regularization—are often introduced as auxiliary loss terms but remain loosely integrated with the architectural design of transformer-based models [12]–[14].

Motivated by these observations, we propose CTCF, a hybrid transformer-based deformable registration framework that integrates a deformable cross-attention encoder, a super-resolution (SR) decoder, and a cascaded refinement strategy within a unified architecture. The proposed design jointly addresses global correspondence modeling and high-resolution deformation reconstruction, while enforcing multiple complementary consistency constraints during training.

Specifically, CTCF combines a deformable cross-attention transformer encoder with a multi-scale SR decoder and progressively refines the deformation field through a cascade of registration stages. Inverse consistency, cycle consistency, and Jacobian-based regularization are jointly incorporated to encourage symmetric, topology-preserving transformations without explicit diffeomorphic parameterization.

We evaluate the proposed method on the OASIS brain MRI benchmark dataset [15] and compare it with state-of-the-art transformer-based deformable registration methods. Experimental results show that CTCF improves deformation regularity and reduces folding artifacts while maintaining competitive registration accuracy.

The main contributions of this work are summarized as follows:

- 1) *A hybrid deformable registration architecture that integrates deformable cross-attention and SR decoding.*
- 2) *A cascaded refinement strategy for progressive deformation improvement.*

3) *A unified training framework* combining inverse consistency, cycle consistency, and topology-preserving regularization.

4) *A systematic experimental analysis* revealing how architectural choices and consistency-based losses affect the trade-off between overlap accuracy and deformation regularity in transformer-based registration.

II. RELATED WORK

A. Learning-based Deformable Registration

Deformable medical image registration has a long research history, comprehensively reviewed in classical surveys [1], [16]. Traditional approaches relied on iterative optimization and parametric deformation models, such as free-form deformations (FFD) based on B-splines or diffeomorphic formulations [17], [18]. While these methods provide strong theoretical guarantees, they are computationally expensive and often unsuitable for large-scale or time-critical applications.

The introduction of deep learning significantly changed this landscape. Balakrishnan et al. proposed one of the first unsupervised convolutional frameworks for deformable registration, directly predicting dense displacement fields from image pairs [3]. This idea was further developed in VoxelMorph [4], which popularized a U-Net-based architecture trained with image similarity and smoothness regularization. VoxelMorph demonstrated that learning-based registration can achieve competitive accuracy while enabling substantially faster inference compared to classical iterative methods.

Subsequent extensions explored probabilistic and diffeomorphic formulations to improve deformation regularity and uncertainty modeling. Despite these advances, purely convolutional approaches remain limited by their local receptive field, which restricts their ability to capture long-range anatomical correspondences, particularly under large or global deformations.

B. Transformer-based Registration Models

To overcome the locality limitations of convolutional networks, transformer architectures were introduced into medical image registration. TransMorph [7] was among the first methods to adapt vision transformers to unsupervised deformable registration. By leveraging window-based self-attention, TransMorph enabled modeling of long-range spatial dependencies while remaining computationally feasible for 3D medical volumes. Compared to convolutional baselines such as VoxelMorph [4], TransMorph demonstrated improved registration accuracy on benchmark brain MRI datasets, including OASIS and IXI.

Follow-up work explored alternative transformer designs for volumetric registration. ViT-V-Net [12] replaced convolutional encoders with global self-attention blocks, further highlighting the potential of transformers for capturing global anatomical context. These approaches build upon general advances in transformer architectures for vision tasks.

However, applying dense attention to high-resolution 3D volumes remains computationally challenging. To address this, recent methods introduced more structured or sparse attention mechanisms. In particular, TransMorph-DCA [19] builds on the TransMorph framework by incorporating deformable cross-attention, enabling sparse and adaptive

sampling of anatomically relevant regions while maintaining scalability. This design substantially improves robustness to large and non-uniform deformations, making TransMorph-DCA a strong transformer-based encoder for deformable medical image registration.

Nevertheless, existing transformer-based methods—including TransMorph and its variants—typically rely on conventional convolutional decoders with interpolation-based upsampling, which may limit the spatial fidelity of the predicted high-resolution deformation fields.

C. Super-Resolution and Hybrid Architectures

Beyond encoder design, another line of work focuses on improving the decoder side of registration networks. Motivated by advances in SR and dense prediction, recent methods treat deformation field reconstruction as a resolution enhancement problem rather than simple interpolation.

UTSRMorph [20] adopts this perspective by formulating deformation field reconstruction as a SR problem. Instead of relying on trilinear interpolation, UTSRMorph employs learned SR blocks to progressively reconstruct high-resolution displacement fields from coarse representations. This design improves deformation smoothness and spatial fidelity, particularly in anatomically complex regions.

In addition, UTSRMorph integrates channel attention mechanisms within the decoder to enhance feature refinement during upsampling. Experimental results reported in [20] demonstrate improved deformation quality and competitive registration accuracy compared to interpolation-based transformer models.

However, the encoder in UTSRMorph relies on window-based self-attention rather than deformable cross-attention, which may limit its ability to capture sparse, long-range correspondences between distant anatomical structures.

D. Consistency and Topology-Preserving Constraints

In addition to architectural advances, enforcing physical and topological consistency has been recognized as a critical aspect of deformable registration. Inverse consistency was introduced as an explicit training objective in inverse-consistent networks [14], demonstrating that enforcing symmetry between forward and backward transformations improves deformation regularity without explicit diffeomorphic parameterization.

CycleMorph [13] further incorporated cycle-consistency at the image level, requiring that forward-backward warping reconstructs the original image. This constraint was shown to reduce folding artifacts and improve topology preservation. Complementary to these approaches, penalties based on the Jacobian determinant of the deformation field are widely used to discourage non-invertible transformations and enforce local diffeomorphism.

Despite their effectiveness, these constraints are often applied in isolation and are rarely integrated into modern transformer-based registration frameworks in a unified manner.

E. Positioning of the Present Work

Existing methods exhibit complementary strengths. Transformer-based models such as TransMorph excel at global correspondence modeling, while convolutional and super-resolution decoders improve high-resolution deformation reconstruction. Consistency-based constraints—

such as inverse consistency, cycle consistency, and Jacobian regularization—provide strong guarantees of deformation regularity but are not fully exploited within hybrid architectures.

The present work bridges these directions by combining a deformable cross-attention transformer encoder, a super-resolution decoder, and a cascaded refinement strategy, while jointly enforcing inverse consistency, cycle consistency, and topology-preserving regularization. This integration directly addresses the limitations of prior approaches and forms the foundation of the proposed CTCF framework.

III. METHODS

A. CTCF Architecture Overview

CTCF is a cascaded encoder-decoder framework for unsupervised 3D deformable medical image registration.

Given a moving image I_M and a fixed image I_F , defined on a discrete spatial domain $\Omega \subset \mathbb{Z}^3$ corresponding to the 3D image grid, the model predicts a dense displacement field Φ :

$\Omega \rightarrow \mathbb{R}^3$, which spatially aligns I_M to I_F . The design of CTCF addresses several key challenges commonly observed in learning-based registration methods, including the modeling of long-range anatomical correspondences, the reconstruction of high-resolution and detail-preserving deformation fields, and the enforcement of physically plausible, topology-preserving transformations.

To achieve this, CTCF integrates a deformable cross-attention (DCA) transformer encoder, a super-resolution (SR) decoder, and a cascaded refinement strategy within a unified architecture (Fig. 1).

B. Deformable Cross-Attention Transformer Encoder

CTCF adopts the Swin Transformer encoder with deformable cross-attention from TransMorph-DCA. The moving and fixed images are first embedded into patch representations and processed through multiple hierarchical transformer stages.

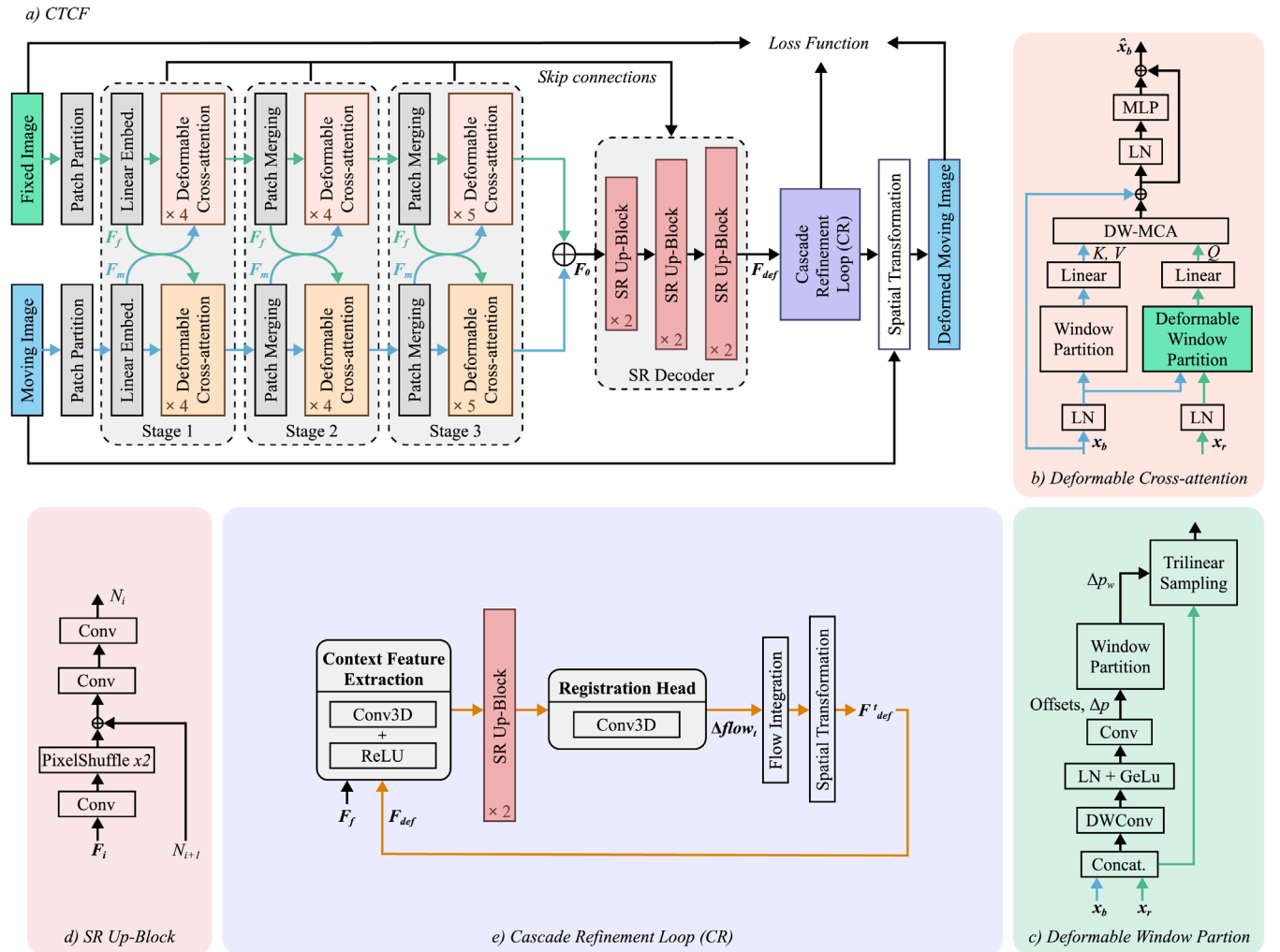


Fig. 1. Overview of the proposed CTCF framework. (a) Overall architecture of the CTCF model, consisting of parallel transformer encoders with deformable cross-attention (DCA), a super-resolution (SR) decoder, and a cascaded refinement loop (CR). The moving and fixed images are processed by symmetric Swin Transformer encoders with DCA blocks, whose multi-scale features are fused and progressively upsampled by SR decoder blocks to generate high-resolution deformation features. (b) Deformable cross-attention (DCA) module, which adaptively aligns features between the moving and fixed image streams using deformable window-based attention. (c) Deformable window partition strategy, where learnable offsets enable sparse and flexible sampling across spatial windows. (d) Super-resolution upsampling (SR) block, which reconstructs higher-resolution feature maps using PixelShuffle-based upsampling and convolutional refinement with optional skip connections. (e) Cascaded refinement loop (CR), where the deformation field is iteratively refined through context feature extraction, super-resolution decoding, flow integration, and spatial transformation, progressively improving alignment accuracy.

Within each stage, window-based self-attention models intra-image contextual relationships, while deformable cross-attention selectively exchanges information between the moving and fixed feature streams. Instead of dense cross-attention, DCA learns a sparse set of sampling offsets, allowing the model to attend to anatomically relevant regions across large spatial distances with reduced computational cost.

This design improves robustness to large and spatially heterogeneous deformations. The encoder produces multi-scale feature maps $\{f_l^M, f_l^F\}$ at different resolutions. At each level l , features from both streams are fused by element-wise summation, which are then forwarded to the decoder:

$$f_l = f_l^M + f_l^F \quad (1)$$

C. Super-Resolution Decoder

To overcome the limitations of interpolation-based upsampling, CTCF employs a super-resolution decoder inspired by UTSRMorph. Instead of trilinear interpolation, the decoder reconstructs high-resolution deformation features using learned upsampling modules.

Each SR block increases spatial resolution by a factor of two via 3D sub-pixel convolution (PixelShuffle), followed by convolutional refinement layers. Skip connections from the encoder to the decoder are enabled in the proposed architecture and are consistently used in all experiments to preserve fine-grained spatial information during upsampling. Both transformer-level and convolution-level skip connections are incorporated, allowing multi-scale features from the encoder to be directly propagated to the decoder and improving the fidelity of high-resolution deformation reconstruction.

Starting from the coarsest encoder level, the decoder progressively upsamples and refines feature representations until full resolution is reached. This strategy enables accurate reconstruction of smooth yet detailed displacement fields and mitigates aliasing artifacts commonly introduced by interpolation-based decoders.

D. Cascaded Refinement Strategy

CTCF incorporates an internal cascade of refinement steps to iteratively improve registration accuracy. Rather than predicting the full deformation field in a single pass, the model estimates a sequence of incremental deformation updates $\Delta\phi_t$.

At cascade step t , the current deformation estimate Φ_t is applied to warp the moving image, yielding an intermediate warped image $I_M^{(t)}$. A lightweight convolutional module extracts appearance discrepancies between $I_M^{(t)}$ and I_F , which are processed by SR-based refinement heads to predict the next update $\Delta\phi_t$.

The deformation is updated via differentiable flow composition:

$$\Phi_{t+1} = \Phi_t + (\Delta\phi_t \circ \Phi_t) \quad (2)$$

This coarse-to-fine refinement process allows CTCF to handle large deformations more effectively and improves convergence stability. In our experiments, we use $T = 12$ cascade steps, matching the configuration of TransMorph-DCA for fair comparison.

E. Loss Functions

CTCF is trained in an unsupervised manner using a composite objective, where all loss terms are computed symmetrically for forward and backward registrations:

$$L = \omega_{sim}L_{sim} + \omega_{smooth}L_{smooth} + \omega_{icon}L_{icon} + \omega_{cyc}L_{cyc} + \omega_{jac}L_{jac}, \quad (3)$$

where L_{sim} denotes the image similarity loss implemented as local normalized cross-correlation (NCC), L_{smooth} is a gradient-based regularization term encouraging spatial smoothness of the displacement field, L_{icon} enforces inverse consistency between forward and backward transformations, L_{cyc} denotes the image-level cycle-consistency loss, and L_{jac} penalizes non-positive Jacobian determinants to discourage local foldings. The corresponding loss weights are fixed across all experiments and set to $\omega_{sim} = 1.0$, $\omega_{smooth} = 1.0$, $\omega_{icon} = 0.1$, $\omega_{cyc} = 0.05$, $\omega_{jac} = 0.01$.

1) *Image Similarity Loss* measures similarity between the warped moving image and the fixed image.

2) *Smoothness Regularization* discourages abrupt spatial variations in the displacement field.

3) *Inverse-Consistency Loss* reduces asymmetry between forward and backward transformations.

4) *Cycle-Consistency Loss* enforces reconstruction of the original image after forward-backward warping.

5) *Jacobian Determinant Penalty* suppresses local foldings by penalizing non-invertible deformations.

For baseline methods, only the loss terms originally defined in the respective implementations are used. In particular, TransMorph-DCA and UTSRMorph do not incorporate inverse-consistency, cycle-consistency, or Jacobian-based penalties unless explicitly stated.

IV. EXPERIMENTS

A. Dataset and Preprocessing

All experiments are conducted on the OASIS T1-weighted brain MRI dataset. The dataset contains 413 3D volumes with anatomical segmentations. We follow the commonly used protocol adopted in recent transformer-based registration studies by splitting the data into 394 training volumes and 19 validation volumes. During training, moving-fixed pairs are formed as unordered pairs sampled from the training set, and no subject identity information is used for pair construction. The validation subset is used exclusively for inference and performance evaluation.

All volumes are preprocessed to a fixed spatial size of $160 \times 192 \times 224$, and intensities are normalized per volume. To reduce GPU memory consumption, TransMorph-DCA and CTCF are trained with inputs downsampled by a factor of 2 using average pooling, while evaluation is performed at full resolution. Anatomical segmentations are used exclusively for quantitative evaluation. During training, optimization is driven by image similarity and regularization terms; Dice scores are computed only for validation and reporting purposes.

B. Baseline Methods

CTCF is compared against two recent transformer-based deformable registration baselines that represent complementary design choices in modern registration architectures:

1) *TransMorph-DCA* extends the original TransMorph framework by introducing deformable cross-attention, enabling sparse and adaptive modeling of long-range anatomical correspondences between moving and fixed images. This method serves as a strong baseline for evaluating the effectiveness of transformer-based correspondence modeling.

2) *UTSRMorph*, which formulates displacement field prediction as a super-resolution problem and replaces interpolation-based decoding with learned super-resolution modules. It represents a state-of-the-art approach focused on improving high-resolution deformation reconstruction through decoder-side architectural design.

All methods are trained and evaluated using the same data split and pipeline to ensure a fair comparison.

C. Evaluation Metrics

Registration performance is quantitatively assessed using the following metrics:

1) *Dice Similarity Coefficient (DSC)*, computed between warped moving segmentations and fixed segmentations, averaged over anatomical labels 1–35, measuring anatomical overlap accuracy.

2) *Folding percentage*, defined as the proportion of voxels with non-positive Jacobian determinants of the deformation field, serving as an indicator of topological violations.

3) *95th percentile Hausdorff Distance (HD95)*, computed between warped and fixed segmentations and averaged over anatomical labels 1–35, capturing boundary alignment accuracy and sensitivity to local errors.

4) *Deformation regularity*, measured as the standard deviation of the log-Jacobian determinant of the deformation field, where the Jacobian determinant is clamped to a small positive value to ensure numerical stability.

For methods that predict deformations at reduced resolution, the displacement field is upsampled to full resolution using trilinear interpolation with magnitude scaling consistent with the downsampling factor.

D. Implementation details

All models are implemented in PyTorch and trained for 500 epochs using AdamW with a learning rate of 1×10^{-4} and batch size 1, without learning-rate scheduling. TransMorph-DCA and CTCF perform downsampling by average pooling ($\times 2$) during training. Automatic mixed precision is used to reduce memory consumption.

CTCF is trained bidirectionally by computing forward and backward registrations within each iteration to enable inverse- and cycle-consistency constraints. In our setup, TransMorph-DCA and CTCF predict flows at half resolution and upsample them by $\times 2$ with displacement scaling, while UTSRMorph predicts the full-resolution flow in a single forward pass.

Training was performed on an NVIDIA RTX PRO 6000 Blackwell GPU, while inference-time runtime measurements were collected on an NVIDIA RTX 5070 under identical input resolution and preprocessing. Registration accuracy is

evaluated using DSC and HD95, and topology preservation is assessed via folding percentage computed from the Jacobian determinant.

V. RESULTS

A. Quantitative Registration Accuracy

Table I summarizes quantitative registration performance on the OASIS evaluation set in terms of Dice similarity coefficient (DSC), 95th percentile Hausdorff distance (HD95), folding percentage, and deformation regularity measured by the standard deviation of the log-Jacobian determinant. The log-Jacobian statistic is computed after clamping the Jacobian determinant to a small positive value, while topological violations are quantified separately via folding percentage. Arrows indicate whether higher (\uparrow) or lower (\downarrow) values correspond to better performance.

Among the compared methods, TransMorph-DCA achieves the highest Dice score (0.869 ± 0.015), confirming the effectiveness of deformable cross-attention for modeling long-range anatomical correspondences in brain MRI. This strong overlap accuracy is accompanied by a moderate folding rate (0.104%) and a relatively high log-Jacobian variability ($\text{std} = 0.773$), indicating the presence of local deformation irregularities despite good anatomical alignment.

UTSRMorph exhibits substantially lower Dice performance (0.817 ± 0.023) and the highest HD95 value (1.89), reflecting reduced overlap accuracy and inferior boundary alignment. Moreover, UTSRMorph produces deformation fields with a markedly high folding percentage (0.713%) and the largest log-Jacobian variance ($\text{std} = 1.778$), indicating limited topological robustness in its current configuration.

The proposed CTCF achieves a Dice score of 0.859 ± 0.014 , which is lower than TransMorph-DCA but clearly higher than UTSRMorph. Importantly, this moderate reduction in overlap accuracy is accompanied by a substantial improvement in deformation quality. In particular, CTCF reduces the folding percentage by more than $2\times$ compared to TransMorph-DCA and by more than an order of magnitude compared to UTSRMorph, achieving the lowest folding rate across all methods (0.047%). In addition, CTCF yields the lowest log-Jacobian standard deviation (0.571), indicating significantly smoother and more physically plausible deformation fields.

Overall, these results demonstrate that CTCF explicitly trades a small amount of overlap accuracy for a pronounced gain in deformation regularity and topology preservation.

TABLE I. QUANTITATIVE COMPARISON ON OASIS TEST SET

Method	Dice (mean \pm std) \uparrow	HD95 (mean) \downarrow	Fold % (mean) \downarrow	SDlogJ \downarrow
CTCF	0.859 ± 0.014	1.53	0.047	0.571
UTSRMorph	0.817 ± 0.023	1.89	0.713	1.778
TM-DCA	0.869 ± 0.015	1.42	0.104	0.773

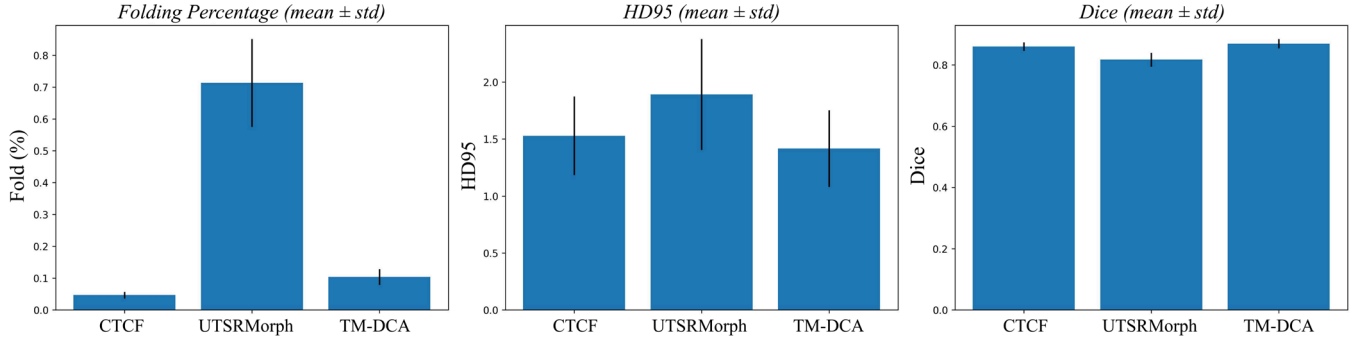


Fig. 2. Quantitative comparison of registration performance on OASIS evaluation set. Bar charts show mean Dice similarity coefficient, HD95, and folding percentage for all compared methods.

B. Topology Preservation Analysis

Beyond overlap accuracy, preserving anatomical topology is a critical requirement for deformable image registration. Folding percentage, derived from the Jacobian determinant of the deformation field, directly reflects the presence of non-invertible transformations.

CTCF consistently exhibits the lowest folding rate (0.047%) on the OASIS evaluation set, in contrast to TransMorph-DCA (0.104%) and UTSRMorph (0.713%). Notably, this behavior is achieved without explicit diffeomorphic parameterization or post-processing, relying solely on training-time regularization.

The joint use of inverse-consistency, cycle-consistency, and Jacobian-based penalties effectively constrains the deformation field during training. The reduced standard deviation of the log-Jacobian determinant further indicates that CTCF produces smoother and more physically plausible deformations than both baseline methods, consistent with the quantitative results in Table I.

C. Qualitative Evaluation

Figure 3 presents qualitative registration results on the OASIS dataset. For each method, deformed grids are visualized on orthogonal anatomical planes (axial, coronal, and sagittal) to assess deformation smoothness and topological behavior.

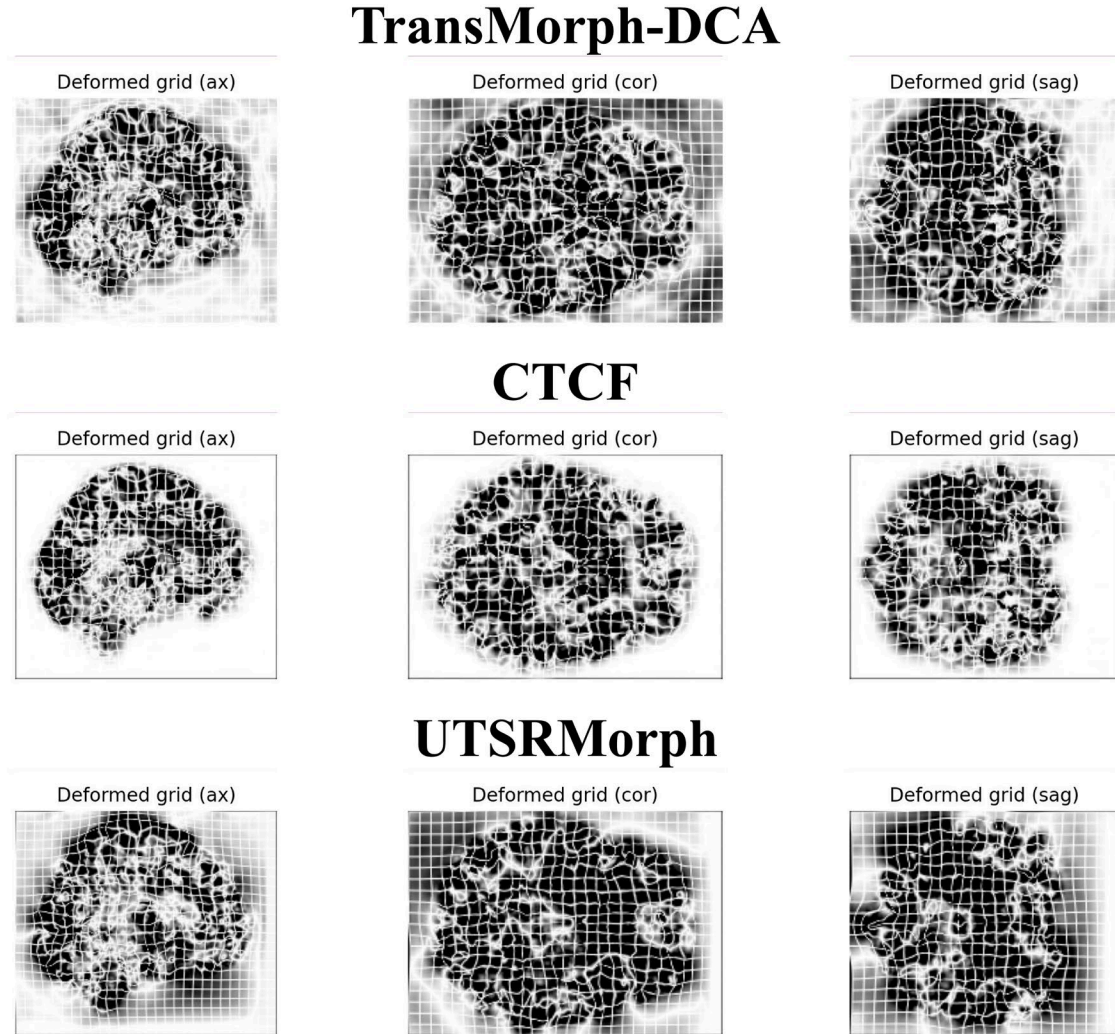


Fig. 3. Qualitative comparison of deformation fields on the OASIS dataset. Deformed grids are shown on axial, coronal, and sagittal planes to illustrate deformation smoothness and topology preservation.

TransMorph-DCA produces accurate global alignment but exhibits localized grid compression and irregular spacing, particularly in regions with complex cortical folding. UTSRMorph shows pronounced grid folding and irregular deformation patterns, consistent with its high folding percentage and log-Jacobian variability.

In contrast, CTCF yields substantially smoother and more regular deformation grids across all views, with minimal grid folding and reduced local distortions. These qualitative observations closely align with the quantitative improvements in folding percentage and deformation regularity, confirming the effectiveness of the proposed consistency-driven cascaded refinement strategy.

D. Computational Efficiency

Inference efficiency was evaluated by measuring the runtime per image pair on the OASIS evaluation set at full spatial resolution under identical inference conditions on an NVIDIA RTX 5070 GPU. TransMorph-DCA achieves the fastest inference among the compared methods with an average runtime of 0.82 s per image pair, reflecting its comparatively lightweight single-pass inference pipeline. UTSRMorph is slower, requiring 1.28 s per pair, which is consistent with the additional compute introduced by learned super-resolution decoding. CTCF incurs the highest inference cost at 1.50 s per pair due to the combined overhead of super-resolution decoding and cascaded refinement. Despite this increase in runtime, CTCF remains practically usable for 3D brain MRI inference. The additional computational cost is accompanied by substantially improved deformation regularity and topology preservation, as evidenced by the lowest folding rate and reduced log-Jacobian variability among all compared methods.

VI. DISCUSSION

Our results clarify how encoder choice, decoder design, and consistency/topology constraints influence accuracy vs. deformation quality in transformer-based 3D brain MRI registration.

1) *Deformable cross-attention is the main driver of overlap accuracy.* TransMorph-DCA achieves the highest mean Dice score on OASIS (0.869), confirming that sparse deformable sampling in attention space is effective for modeling long-range anatomical correspondences. This observation supports the decision to retain a DCA-based transformer encoder as the backbone of CTCF rather than replacing it.

2) *Super-resolution decoding consistency regularization mainly improve deformation plausibility rather than Dice.* CTCF achieves slightly lower Dice compared to TransMorph-DCA (0.859 vs. 0.869), but produces substantially more regular deformation fields. In particular, the folding rate is reduced from 0.104% (TM-DCA) to 0.047% (CTCF), and the standard deviation of the log-Jacobian determinant decreases from 0.773 to 0.571. This indicates that learned upsampling combined with consistency- and topology-aware losses helps preserve fine-scale displacement structure that is otherwise degraded by interpolation-based decoding, yielding smoother and more physically plausible transformations.

3) *Cascaded refinement amplifies the effect of consistency constraints.* The multi-step cascade allows inverse-consistency, cycle-consistency, and Jacobian-based

penalties to act repeatedly on intermediate deformation estimates. This incremental refinement likely suppresses local inconsistencies before they accumulate into foldings. Notably, CTCF achieves near-zero folding without explicit velocity-field integration or diffeomorphic parameterization, relying instead on training-time regularization.

4) *Super-resolution decoding alone is insufficient for topology preservation.* Despite using learned SR decoding, UTSRMorph exhibits inferior deformation quality on OASIS, with a high folding percentage (0.713%) and large log-Jacobian variance (1.778), alongside reduced Dice (0.817) and higher HD95 (1.89). These results suggest that decoder-side super-resolution, when not supported by a strong long-range correspondence encoder and explicit consistency/topology constraints, does not guarantee anatomically plausible deformations.

While CTCF substantially improves deformation regularity, it incurs higher computational cost due to the combination of super-resolution decoding and cascaded refinement. Inference time increases from 0.82 s per image pair for TransMorph-DCA to 1.50 s for CTCF, compared to 1.28 s for UTSRMorph. This overhead reflects the architectural complexity of the proposed model and motivates future work on reducing cascade depth or introducing adaptive refinement strategies to balance efficiency and topology preservation.

VII. CONCLUSION

We presented CTCF, a cascaded unsupervised 3D deformable registration framework that integrates a deformable cross-attention Swin transformer encoder, a super-resolution decoder for high-resolution displacement reconstruction, and unified consistency- and topology-aware regularization.

Experiments on the OASIS dataset demonstrate that CTCF achieves the intended trade-off between accuracy and deformation quality. While TransMorph-DCA attains the highest Dice score (0.869), CTCF maintains competitive overlap accuracy (0.859) while substantially improving topological correctness, reducing folding rates (0.047% vs. 0.104%) and deformation irregularity (Std(log J) 0.571 vs. 0.773). These results show that physically plausible deformations can be achieved without explicit diffeomorphic parameterization by combining architectural design with appropriate training-time constraints.

Overall, CTCF illustrates that hybridizing deformable cross-attention encoding with super-resolution decoding and consistency-driven objectives is a promising direction for improving the robustness and reliability of transformer-based medical image registration. Future work will focus on reducing computational overhead, extending evaluation to more anatomically diverse datasets, and exploring adaptive cascade strategies to further improve efficiency without sacrificing topological guarantees.

REFERENCES

- [1] A. Sotiras, C. Davatzikos, and N. Paragios, “Deformable medical image registration: A survey,” *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013.
- [2] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, “Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain,” *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, Feb. 2008.
- [3] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, “An unsupervised learning model for deformable medical image

- registration,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9252–9260.
- [4] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, “VoxelMorph: A learning framework for deformable medical image registration,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.
 - [5] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, “Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces,” *Medical Image Analysis*, vol. 57, pp. 226–236, Oct. 2019.
 - [6] A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
 - [7] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du, “TransMorph: Transformer for unsupervised medical image registration,” *Medical Image Analysis*, vol. 82, Art. no. 102615, Nov. 2022.
 - [8] J. Chen, E. C. Frey, and Y. Du, “Unsupervised learning of diffeomorphic image registration via TransMorph,” in *Biomedical Image Registration (WBIR)*, Springer, 2022, pp. 96–102.
 - [9] N. A. Nefediev, N. E. Staroverov, and R. V. Davydov, “Improving compliance of brain MRI studies with the atlas using a modified TransMorph neural network,” *St. Petersburg State Polytechnical University Journal. Physics and Mathematics*, vol. 17, no. 3, pp. 335–339, 2024, doi: 10.18721/JPM.173.168.
 - [10] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, 2020.
 - [11] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 9992–10002.
 - [12] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, “ViT-V-Net: Vision transformer for unsupervised volumetric medical image registration,” in *Medical Imaging with Deep Learning (MIDL)*, 2021.
 - [13] B. Kim, D. H. Kim, S. H. Park, J. Kim, J.-G. Lee, and J. C. Ye, “CycleMorph: Cycle consistent unsupervised deformable image registration,” *Medical Image Analysis*, vol. 71, Art. no. 102036, Jul. 2021.
 - [14] Y. Zhang, “Inverse-consistent deep networks for unsupervised deformable image registration,” arXiv preprint arXiv:1809.03443, 2018.
 - [15] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, “Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle-aged, nondemented, and demented older adults,” *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007.
 - [16] D. L. G. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, “Medical image registration,” *Physics in Medicine & Biology*, vol. 46, no. 3, pp. R1–R45, Mar. 2001.
 - [17] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, “Nonrigid registration using free-form deformations: Application to breast MR images,” *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, Aug. 1999.
 - [18] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, “Diffeomorphic demons: Efficient non-parametric image registration,” *NeuroImage*, vol. 45, no. 1, pp. S61–S72, Mar. 2009.
 - [19] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, “TransMorph-DCA: Deformable cross-attention for transformer-based medical image registration,” arXiv preprint arXiv:2303.06179, 2023.
 - [20] Y. Zheng, Z. Wang, B. Huang, N. H. Lim, and B. W. Papież, “UTSRMorph: Unified Transformer and Super-Resolution Framework for Unsupervised Deformable Medical Image Registration,” *IEEE Transactions on Medical Imaging*, vol. 44, no. 2, pp. 902–916, Feb. 2025.