CONTRAST TRANSFORMATION-GUIDED MULTIMODAL MEDICAL IMAGE REGISTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal medical image registration, especially MRI/CT, is challenging due to inconsistent appearance and large nonlinear deformations. Methods based on hand-crafted metrics are noise-sensitive and require expensive iterative optimization, while deep models still struggle to separate modality-induced intensity differences from true misalignment. Recent translation-based methods normalize appearance before registration, but weak coupling between synthesis and alignment and possible geometric distortion can introduce spurious deformations. We propose an end-to-end framework with a contrast transformation branch that aligns gray-level distributions, mapping CT to MR-like contrast without changing anatomy, directly guided by registration loss. On two public CT-MR datasets, our approach achieves state-of-the-art performance, surpassing existing methods by 7.34% and 4.29%. The proposed contrast transformation network improves seven established models, with DSC gains up to 11.9% and 8.0%. These results demonstrate state-of-the-art accuracy, strong generalization, and easy integration into dual-stream registration frameworks.

1 INTRODUCTION

Multi-modal medical image registration establishes spatial correspondences across imaging modalities and underpins many clinical and research tasks (Choo et al., 2025; Hu et al., 2025; Guo et al., 2026; Zhu et al., 2025). Yet accurate registration, especially for MRI/CT, remains difficult due to substantial differences in imaging physics and contrast mechanisms, as well as complex non-linear anatomical deformations (Song et al., 2025; Zubair et al., 2025; Fu et al., 2026).

Multi-modal registration faces two core challenges (Zhu et al., 2025): (i) *cross-modal appearance inconsistency*—the same anatomy exhibits different intensity distributions and contrasts across modalities; and (ii) *large non-linear deformations*—physiological motion and patient repositioning cause complex spatial misalignments. Together, these factors break the intensity-matching assumption, rendering simple intensity-based similarity metrics insufficient for robust registration.

Early methods reduced appearance discrepancies using hand-crafted cross-modal similarity metrics, such as mutual information (Guo, 2024) and the modality independent neighborhood descriptor (MIND) (Heinrich et al., 2012). However, these metrics can be sensitive to noise, struggle to capture fine-grained anatomy, and require good initialization for large deformations (Reithmeir et al., 2025). Moreover, iterative optimization is computationally expensive and susceptible to local minima in complex anatomy.

Learning-based registration predicts deformation fields directly from image pairs and substantially reduces the computational cost of pair-specific optimization. Since the introduction of unsupervised deformation prediction frameworks such as VoxelMorph (Balakrishnan et al., 2019), increasingly expressive architectures have been developed to model spatial correspondence. Later methods incorporated Transformers for long-range dependencies (Chang et al., 2024; Chen et al., 2022), cross-modal attention for improved correspondence modeling (Song et al., 2022), and multi-scale pyramid networks for large deformations (Yuan et al., 2025; Wang et al., 2024). However, learning directly from raw multi-modal inputs poses a fundamental dilemma: the model must disentangle modality-specific appearance differences from true spatial misalignment, while intensity differences can reflect either modality shifts or misregistration, hindering accurate correspondence learning (Hu et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A natural strategy is therefore to reduce the modality gap before estimating spatial correspondence (Zhou et al., 2024; Greer et al., 2025). Translation-based registration (Wei et al., 2025; Chen et al., 2025) follows a "unify appearance, then estimate deformation" strategy: an image-to-image translation network maps one modality to the appearance domain of another, turning the task into an approximately mono-modal problem and enabling direct intensity-based losses (Kim et al., 2025; Luo et al., 2026). Representative methods, including RegGAN (Kong et al., 2021), INNReg (Guo, 2024), and AC-CAR (Wang et al., 2026), have explored this paradigm from different perspectives, ranging from registration-aware translation and geometry-conscious modality conversion to contrast-adaptive correspondence learning. Collectively, these studies demonstrate that reducing appearance heterogeneity can simplify correspondence estimation and improve the applicability of intensity- or feature-based registration objectives. However, they also reveal a fundamental trade-off: a stronger appearance transformation may facilitate cross-modal matching, but it may simultaneously alter spatially important structures or introduce appearance patterns that do not faithfully correspond to the original anatomy. Thus, the key challenge is not merely to make the two modalities visually similar, but to identify and transform only the modality-dependent contrast information while leaving the spatial content required for registration unchanged.

We propose a novel registration framework with a contrast transformation branch that maps CT images to an MR-like contrast space by aligning their gray-level distributions, thereby facilitating cross-modal similarity measurement. Unlike full image synthesis, the proposed transformation focuses on adjusting image contrast while explicitly encouraging the preservation of the spatial anatomy, reducing the risk of translation-induced geometric distortions. The transformed CT and MR images are subsequently processed by a shared-weight encoder. Because their gray-level distributions have been brought into a more comparable space, the shared encoder is discouraged from encoding modality-specific intensity patterns and is instead encouraged to capture anatomical structures that remain consistent across modalities. This design separates appearance normalization from spatial correspondence estimation, reducing the burden on the registration network to disentangle modality variation from geometric misalignment. The framework is trained end-to-end, allowing the registration objective to directly guide the contrast transformation toward representations that benefit correspondence estimation rather than visual realism. Experiments on two challenging multi-modal benchmarks demonstrate competitive registration accuracy and consistent improvements when the proposed module is integrated into a range of representative dual-stream registration frameworks.

2 Related Work

Multi-modal Medical Image Registration. Multi-modal medical image registration aims to establish dense anatomical correspondence between images acquired using different imaging mechanisms (Song et al., 2021). Conventional approaches typically optimize statistical similarity measures, such as mutual information, or modality-independent structural descriptors. In particular, the modality-independent neighborhood descriptor (MIND) represents local self-similarity patterns and has been widely adopted for CT–MR registration (Heinrich et al., 2012). Nevertheless, statistical measures can be sensitive to local intensity variations, while hand-crafted descriptors may be insufficiently discriminative for anatomically complex regions.

Recent learning-based methods replace iterative pairwise optimization with direct deformation prediction. VoxelMorph established a widely used unsupervised framework for estimating dense deformation fields (Balakrishnan et al., 2019), while coarse-to-fine and recursive architectures improved the recovery of large and complex deformations (Wang et al., 2023; 2024). However, directly applying these general-purpose architectures to multi-modal images is challenging because corresponding tissues may exhibit substantially different intensities, textures, and boundary responses. Recent multi-modal methods therefore explicitly bridge the modality gap during feature extraction, correspondence estimation, or similarity measurement. Cross-modal Attention-Reg combines contrastive representation learning with cross-modal attention to correlate heterogeneous features (Song et al., 2022), whereas modality-agnostic structural representation learning encodes anatomical self-similarity patterns shared across modalities (Mok et al., 2024). ARDMR disentangles modality-specific and modality-invariant representations and employs recursive inference to address large deformations (Hu et al., 2025), while SynMSE learns a cross-modal similarity evaluator that is robust to complex grayscale distribution discrepancies (Zhu et al., 2025). Despite their effectiveness, modality discrepancy and spatial misalignment are generally handled within the same feature-learning

or similarity-estimation pathway, leaving ambiguity as to whether a local discrepancy originates from imaging contrast or anatomical displacement.

Modality-Invariant Representation and Similarity Learning. A major line of research addresses the modality gap by learning contrast-invariant representations or cross-modal similarity functions. Cross-modal Attention-Reg combines contrastive pre-training with cross-modal attention to correlate representations extracted from different modalities (Song et al., 2022). Mok et al. introduced deep neighborhood self-similarity and anatomy-aware contrastive learning to obtain modality-agnostic structural representations without requiring pre-aligned images or anatomical labels (Mok et al., 2024). ARDMR further disentangles modality-specific and modality-invariant representations and employs recursive inference to handle large deformations (Hu et al., 2025).

Instead of explicitly learning a shared representation, SynMSE learns a plug-and-play multi-modal similarity evaluator from simulated spatial misalignments and grayscale distribution variations (Zhu et al., 2025). These approaches reduce the dependence on raw intensity correspondence, but their performance relies on the discriminability and generalizability of the learned representation or similarity evaluator. Moreover, modality discrepancy and spatial correspondence are generally handled within the same optimization pathway, making it difficult to distinguish intensity variations caused by imaging physics from those caused by geometric misalignment.

Translation-Assisted Registration. Another family of methods first converts heterogeneous inputs into a common appearance domain and subsequently performs approximately mono-modal registration. RegGAN incorporates registration consistency into image translation to alleviate the adverse effects of spatially misaligned training pairs (Kong et al., 2021). INNReg employs an invertible translation network to improve geometric consistency during modality conversion and combines the translated images with an unsupervised registration network (Guo, 2024). AC-CAR adopts adaptive conditional modeling to learn registration across varying imaging contrasts (Wang et al., 2026). Related studies have also explored synthesis-guided similarity learning and intensity regression to reduce cross-modal appearance discrepancies.

However, conventional image-to-image translation is commonly optimized for visual realism, reconstruction, or cycle consistency, none of which strictly guarantees spatial invariance. A translator may therefore modify organ boundaries or fine anatomical structures, causing the subsequent registration network to compensate for synthesis-induced distortions rather than the actual inter-image deformation.

Unlike existing translation- or feature-based approaches, our method introduces an auxiliary contrast transformation branch that maps the moving CT toward an MR-like gray-level distribution without pursuing photorealistic image synthesis. The original CT, transformed CT, and MR images are processed by a shared-weight encoder and jointly optimized for registration, encouraging the encoder to emphasize anatomical structures while suppressing modality-specific intensity variations. The resulting module reduces the modality gap at both the image and feature levels and can be integrated into different encoder-decoder registration networks without modifying their deformation decoders.

3 METHODS

Motivated by the aforementioned challenges, we present a registration framework that leverages contrast transformation for improved structural alignment, enabling robust multi-modal registration. We first summarize the overall architecture and then detail each component.

Fig. 1 presents an overview of our framework, which comprises two principal components: 1) a dual-stream pyramid registration network designed for multi-modal image registration, and 2) a contrast transformation network that performs contrast conversion for multi-modal images.

Dual-Stream Feature Pyramid Encoder. We employ a dual-stream feature encoder with shared weights to independently extract feature pyramid maps from the MR image I_{MR} and the CT image I_{CT}, respectively (Huang & Sui, 2025). The encoder comprises four convolutional blocks, each with a convolutional layer followed by InstanceNorm. The first block uses 8 kernels; subsequent blocks double the kernel count while halving spatial resolution to capture multi-scale features. The

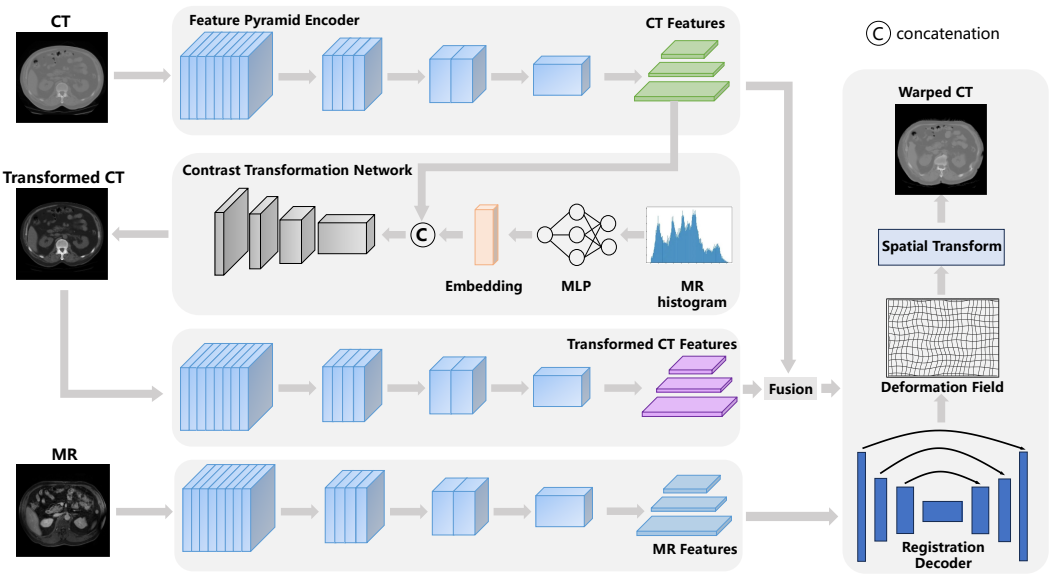

Figure 1: Overview of the proposed framework. The original CT, transformed CT, and MR images are processed through three forward passes of the same shared-weight feature pyramid encoder. Their multi-scale features are fused by the registration decoder to estimate the CT-to-MR deformation field. The three encoder paths are displayed separately for clarity but share all parameters.

resulting feature maps corresponding to the CT and MRI modalities are denoted as *CT Features* and *MR Features*, respectively (as shown in Fig. 1).

Contrast Transformation Network. We propose a unified multi-modal registration framework that jointly performs contrast transformation (Avi-Aharon et al., 2023) and anatomical alignment within a shared latent representation space. As shown in Fig. 1, given a moving CT image I_{CT} and a fixed MR image I_{MR}, both images are first propagated through a shared encoder to extract modality-invariant semantic features. The use of shared weights imposes a common latent space, thereby mitigating inter-modality discrepancies at the representation level while preserving the underlying anatomical structures. Subsequently, a contrast transformation network (TransNet) decodes this latent representation to generate an MR-like CT image I_{TransCT}. The decoder comprises four transposed convolutional layers with 128, 64, 32, and 16 kernels, respectively, which progressively upsample the feature maps by a factor of two at each stage. A final 1×1 convolutional layer followed by a Sigmoid activation function produces the output image I_{TransCT}.

We aim to generate a CT image exhibiting MR-like intensity characteristics to facilitate similarity-based assessment. To prevent the network from encoding MRI-specific anatomy that could introduce geometric distortions, the MR image is not used as a direct input to the translation module. Instead, its intensity histogram is encoded by an MLP to obtain a compact representation in the embedding space. Because the histogram describes the global gray-level distribution without retaining the spatial arrangement of image content, this representation provides modality-specific contrast information while avoiding the direct introduction of MR anatomical structures. The resulting MR histogram embedding is fused with the CT features extracted by the shared encoder and subsequently decoded to produce the transformed image I_{TransCT}. Consequently, the contrast characteristics of I_{TransCT} are conditioned on the MR histogram, whereas its spatial structure is primarily determined by the CT features. This design encourages the encoder to emphasize anatomical information in the CT image and reduce its dependence on modality-specific intensity distributions, thereby facilitating subsequent cross-modal correspondence estimation.

To further stabilize the training process and preserve structural fidelity, we introduce a dynamic feature-mixing strategy that integrates the original CT features (F_{CT}) with the transformed CT features (F_{TransCT}) as:

$$F_{\mathrm{MixCT}} = \lambda * F_{\mathrm{CT}} + (1 - \lambda) * F_{\mathrm{TransCT}}, \tag{1}$$

where λ decreases linearly from 1 to 0.1 over the course of training. The lower bound of 0.1 guarantees that a non-negligible proportion of the original CT features is always preserved, thereby mitigating potential structural degradation induced by the transformed CT. This mechanism promotes robust preservation of anatomical structures in the early training stages, while gradually facilitating improved cross-contrast adaptation as training proceeds.

Hierarchical Image Registration Network. The mixed image features, in conjunction with the MR-derived features, are subsequently input into a hierarchical registration decoder that estimates a dense CT-to-MR deformation field ϕ in a coarse-to-fine manner. Initially, a coarse deformation field is inferred from low-resolution feature maps of both inputs. This preliminary estimate is then iteratively refined by incorporating progressively higher-resolution feature maps at successive stages, ultimately yielding a high-precision deformation field. The process can be formulated as:

$$\phi^l = \mathcal{R}^l \left(F^l_{\mathrm{MixCT}}, F^l_{\mathrm{MR}}, \uparrow \phi^{l+1} \right), \quad l = L-1, \ldots, 1, \tag{2}$$

where $\{F^l_{\mathrm{MixCT}}, F^l_{\mathrm{MR}}\}_{l=1}^{L}$ denote the multi-scale feature pyramids, where $l = L$ represents the coarsest resolution. $\mathcal{R}^l$ denotes the registration block at level l, $\uparrow$ denotes spatial upsampling, and ϕ^l is the refined deformation estimate. The finest estimate ϕ^1 is used as the final deformation field ϕ.

Ultimately, the original CT image is warped using the deformation field ϕ to obtain the spatially aligned output:

$$I_{\mathrm{warped}} = I_{\mathrm{CT}} \circ \phi, \tag{3}$$

where $\circ$ denotes the warping operation applied to I_{CT} using ϕ. In this study, we employ a spatial transformer network (STN) (Jaderberg et al., 2015) to implement the operation.

Notably, the registration decoder is task-agnostic and can be straightforwardly substituted with any existing registration model, thereby underscoring the plug-and-play nature of our contrast transformation module.

Loss Function. We employ a composite loss function that jointly optimizes both the registration network and the contrast transformation network. Because the transformed CT image and the fixed MR image have more comparable contrast characteristics, local normalized cross-correlation (LNCC) is used to measure their local correspondence after spatial warping:

$$\mathcal{L}_{\mathrm{lncc}} = 1 - \mathrm{LNCC}\left(I_{\mathrm{TransCT}} \circ \phi, I_{\mathrm{MR}}\right). \tag{4}$$

LNCC is computed over local windows of size $9 \times 9 \times 9$, allowing the objective to account for spatially varying intensity relationships.

We employ Earth Mover's Distance (EMD) to align the gray-level distribution of the transformed CT image with that of the target MR image:

$$\mathcal{L}_{\mathrm{emd}} = \mathrm{EMD}\left(\mathcal{H}(I_{\mathrm{TransCT}}), \mathcal{H}(I_{\mathrm{MR}})\right), \tag{5}$$

where $\mathcal{H}(\cdot)$ denotes a differentiable normalized histogram. This objective transfers global MR contrast statistics to the transformed CT without directly providing the spatial MR image to the transformation decoder.

To retain information from the original CT image during contrast transformation, mutual information is maximized between the transformed and original CT images:

$$\mathcal{L}_{\mathrm{mi}} = -\,\mathrm{MI}\left(I_{\mathrm{TransCT}}, I_{\mathrm{CT}}\right). \tag{6}$$

This term encourages the transformation to preserve statistical dependence on the original CT content while changing its gray-level distribution.

Spatial smoothness is imposed on the predicted displacement field:

$$\mathcal{L}_{\mathrm{smooth}} = \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \|\nabla \phi(\mathbf{x})\|_2^2, \tag{7}$$

where Ω denotes the image domain. The regularizer suppresses abrupt local variations and encourages spatially coherent deformations.

The overall objective is defined as:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{lncc}} + \mathcal{L}_{\mathrm{emd}} + \mathcal{L}_{\mathrm{mi}} + \mathcal{L}_{\mathrm{smooth}}. \tag{8}$$

4 EXPERIMENTAL SETUP

Datasets. We evaluate the proposed method on two abdominal CT–MR registration settings: Learn2Reg CT-MR (Hering et al., 2022) and 3D AbdomenCT-1K (Ma et al., 2021).

The Learn2Reg CT-MR dataset (Hering et al., 2022) contains paired abdominal CT and MR volumes with manual annotations of four organs: liver, spleen, left kidney, and right kidney. All the volumes have been cropped to the size of $144 \times 160 \times 192$ $(2\,\mathrm{mm} \times 2\,\mathrm{mm} \times 2\,\mathrm{mm})$. CT intensities were clipped to $[-800, 800]$ HU, and both CT and MR volumes were independently normalized to $[0, 1]$. We use eight paired CT-MR volumes, of which three are used for training and five for testing. The dataset additionally contains 90 unpaired abdominal CT and MR scans. From these modality-specific pools, we randomly construct 200 unpaired CT-MR image pairs before training. The resulting pairs are kept fixed across all experiments to ensure a consistent comparison.

The 3D AbdomenCT-1K dataset contains 1,000 CT and 50 labeled MR volumes. All volumes are cropped to $160 \times 160 \times 64$ at $1.75 \times 1.75 \times 2$ mm resolution. CT intensities are clipped to $[-400, 400]$ HU, and both modalities are independently normalized to $[0, 1]$. We select 30 annotated volumes per modality, using 24 for training and six for testing, and construct 200 additional fixed training pairs from the remaining 20 MR volumes and CT pool. All CT–MR pairs are unpaired and inter-subject. Anatomical labels are used only for evaluation, with identical splits and pairs across methods.

Evaluation Metrics. We evaluate registration accuracy using the Dice similarity coefficient (DSC), the 95th-percentile Hausdorff distance (HD95), and the average symmetric surface distance (ASSD). DSC measures the volumetric overlap between corresponding anatomical structures, whereas HD95 and ASSD quantify boundary alignment and are reported in millimeters. Higher DSC and lower HD95 and ASSD indicate better registration accuracy. We additionally report the percentage of voxels with non-positive Jacobian determinant (NJD) to assess the regularity and reliability of the predicted deformation field. A lower NJD indicates fewer local foldings and better topology preservation.

Implementation Details. All experiments were conducted using PyTorch on an NVIDIA RTX 4090 GPU (24 GB of memory). No additional random geometric or intensity augmentation was applied. The networks were trained for 100 epochs with a batch size of 1 using the Adam optimizer. The learning rate was initialized to 1×10^{-4} and linearly decayed to 1×10^{-6} after the first 20 epochs. All experiments used a fixed random seed of 42 and no multi-round training was conducted.

5 RESULTS

Comparison with the State-of-the-Art Approaches. We compared our method against six state-of-the-art registration approaches: VoxelMorph (Balakrishnan et al., 2019), RDP (Wang et al., 2024), MambaMorph (Wang et al., 2025), TransMatch (Chen et al., 2023), INNReg (Guo, 2024), and Attention-Reg (Song et al., 2022). These methods cover complementary registration paradigms, including convolutional deformation prediction, coarse-to-fine pyramidal refinement, state-space modeling, explicit feature matching, translation-assisted registration, and cross-modal attention. They therefore provide a diverse comparison in terms of both deformation estimation and cross-modal correspondence modeling. Our method adopts RDP as the backbone network, utilizing the same feature encoder and registration decoder.

As reported in Table 1, our method achieves the best performance across all registration-accuracy metrics on both datasets. On Learn2Reg MR–CT, it obtains a DSC of 0.658, outperforming the second-best result of 0.613 achieved by MambaMorph by 7.34%. It also reduces HD95 from 11.929 to 11.740 and ASSD from 4.620 to 4.324 compared with the best competing results for the respective metrics. On the AbdomenCT-1K-based CT–MR setting, RDP is the strongest baseline, achieving a DSC of 0.559, an HD95 of 11.086, and an ASSD of 3.724. Our framework improves these values to 0.583, 10.192, and 3.646, respectively. This corresponds to a 4.29% relative improvement in DSC and an 8.06% reduction in HD95. These improvements are particularly notable given the unpaired and inter-subject nature of this setting. The relative ranking of the competing methods varies across the two datasets. For example, MambaMorph performs strongly on Learn2Reg MR–CT but exhibits lower accuracy on the AbdomenCT-1K-based setting, whereas RDP remains comparatively robust across both tasks. In contrast, our method consistently achieves the best DSC, HD95, and ASSD. The

Table 1: Quantitative comparison with state-of-the-art methods on two datasets. The best results are highlighted in bold.

Methods	Learn2Reg MR-CT				3D AbdomenCT-1K			
	DSC ↑	HD95 ↓	ASSD ↓	NJD (%) ↓	DSC ↑	HD95 ↓	ASSD ↓	NJD (%) ↓
Initial	0.490	12.680	5.328	–	0.344	13.878	5.703	–
VoxelMorph	0.504	12.474	5.208	0.168	0.483	12.709	4.668	0.286
RDP	0.597	12.099	4.620	0.052	0.559	11.086	3.724	0.041
MambaMorph	0.613	11.929	4.653	0.327	0.480	13.031	4.964	0.612
TransMatch	0.591	12.193	4.762	0.108	0.505	12.864	4.875	0.254
INNReg	0.501	12.553	5.248	0.143	0.354	13.748	5.707	0.074
Attention-Reg	0.518	12.302	5.056	0.217	0.405	13.372	5.843	0.408
Ours	**0.658**	**11.740**	**4.324**	**0.030**	**0.583**	**10.192**	**3.646**	**0.039**

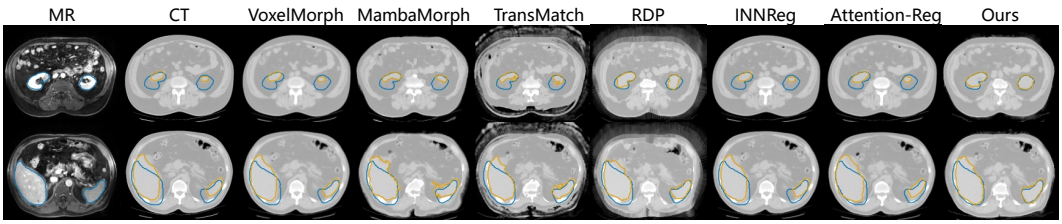

Figure 2: Visualization of registration results obtained by different methods. The organ contours in the reference MR image are depicted with blue lines, whereas the organ contours in the corresponding registered image are shown with orange lines.

NJD value remains comparable to or lower than those of the competing methods, indicating that the improvement in registration accuracy is not achieved at the cost of increased local folding.

Fig. 2 shows that our method achieved more precise alignment at organ boundaries, demonstrating that effective contrast transformation enables more accurate structural correspondence.

Enhancement of Established Registration Models. To evaluate the generalizability of our approach, we incorporated the contrast transformation network into seven representative dual-stream registration models: RDP (Wang et al., 2024), CGNet (Chang et al., 2024), CorrMLP (Meng et al., 2024), ModeT (Wang et al., 2023), PCNet (Lv et al., 2022), PIViT (Ma et al., 2023), and PRNet++ (Kang et al., 2022).

For each baseline, we integrate TransNet as a plug-and-play auxiliary branch without modifying its original registration decoder. The original CT, transformed CT, and MR images are processed by a shared-weight encoder.

As reported in Table 2, our TransNet consistently improved registration accuracy across all seven evaluated models on the Learn2Reg 2022 CT-MR dataset, yielding DSC score gains ranging from 1.59% to 11.9%. Among these, ModeT achieved the highest DSC score of 0.704. The greatest relative gain was observed for PRNet++ with the DSC score increasing from 0.513 to 0.574, corresponding to an 11.9% boost. Furthermore, all models exhibited reductions in HD95, ASSD, and NJD metrics. A similar trend was observed on the 3D AbdomenCT-1K dataset, where incorporating our TransNet yielded performance gains across all baseline models.

Collectively, these results demonstrate that the proposed contrast transformation branch consistently benefits the evaluated registration architectures. By explicitly separating contrast transformation from spatial alignment, TransNet provides a practical encoder-side extension without requiring changes to the original registration decoders. This design reduces the influence of modality-specific appearance variations, allowing the registration networks to focus more on anatomical correspondence.

Table 2: Quantitative results on two datasets. + *TransNet* denotes the inclusion of our proposed contrast transformation method.

Model	Learn2Reg MR-CT				3D AbdomenCT-1K			
	DSC ↑	HD95 ↓	ASSD ↓	NJD (%) ↓	DSC ↑	HD95 ↓	ASSD ↓	NJD (%) ↓
RDP	0.597	12.099	4.620	0.052	0.559	11.086	3.724	0.041
+ TransNet	**0.658**	**11.740**	**4.324**	**0.030**	**0.583**	**10.192**	**3.646**	**0.039**
CGNet	0.521	12.232	4.909	4.458	0.442	11.492	4.447	5.304
+ TransNet	**0.556**	**11.659**	**4.703**	**2.300**	**0.464**	**10.888**	**4.073**	**4.050**
CorrMLP	0.578	12.028	4.715	4.559	0.563	8.815	3.338	3.200
+ TransNet	**0.590**	**11.322**	**4.382**	**4.328**	**0.586**	**8.710**	**3.180**	**0.877**
ModeT	0.688	10.370	3.805	0.055	0.556	9.806	3.819	0.207
+ TransNet	**0.704**	**9.662**	**3.479**	**0.033**	**0.577**	**9.238**	**3.493**	**0.155**
PCNet	0.692	10.005	3.594	1.241	0.515	10.776	4.263	0.072
+ TransNet	**0.703**	**9.740**	**3.485**	**0.981**	**0.532**	**10.701**	**4.251**	**0.037**
PIViT	0.576	11.476	4.542	0.206	0.425	11.202	4.477	0.610
+ TransNet	**0.610**	**11.129**	**4.382**	**0.154**	**0.459**	**10.020**	**3.998**	**0.312**
PRNet++	0.513	12.565	5.152	3.096	0.361	12.621	5.531	1.239
+ TransNet	**0.574**	**12.349**	**5.007**	**3.093**	**0.373**	12.721	**5.483**	**1.216**

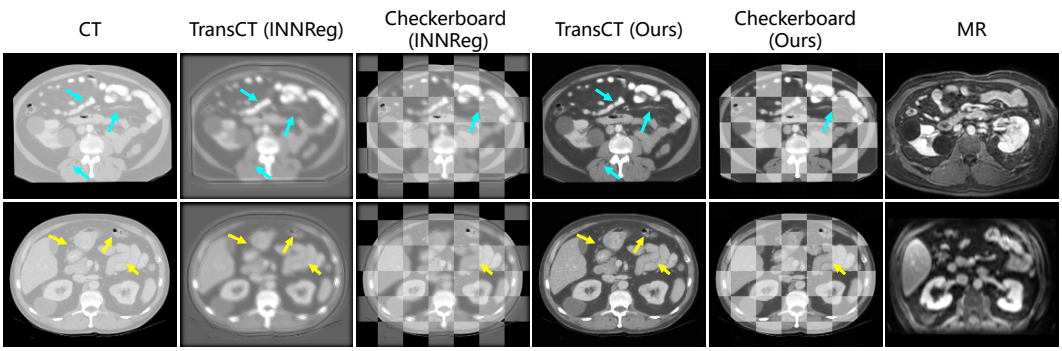

Figure 3: Visual comparison of the original CT, the converted CT (TransCT), the checkerboard fusion image (TransCT & CT), and the original MR. Continuous anatomical structures across the checkerboard squares indicate anatomy preservation, whereas discontinuities reveal structural deformation.

Contrast Transformation with Structural Fidelity. A primary concern in translation-based registration is that generative translators may inadvertently modify anatomy, thereby introducing spurious deformations. To investigate whether our approach mitigates this risk, Fig. 3 depicts the transformed CT images I_{TransCT} generated by our method in comparison with those produced by the INNReg method. These results suggest that our transformation introduces fewer structural misalignments. This observation is further supported by the gradient profiles (as shown in Fig. 4); the principal peaks of our result largely coincide with those of the original CT, whereas INNReg produces several additional high-gradient responses at locations without corresponding peaks in the original image. They are also consistent with the intended role of the shared encoder, which emphasizes anatomical structure while reducing sensitivity to modality-specific contrast variations, thereby providing more reliable features for subsequent registration. By explicitly constraining the contrast transformation to maintain anatomical consistency, the resulting deformation fields predominantly captured true misalignment rather than translation-induced artifacts. Consequently, the multi-modal registration problem reduced to an alignment of structural representations within a shared latent space, thereby overcoming the geometric deformation limitations that characterize existing translation-based registration methods.

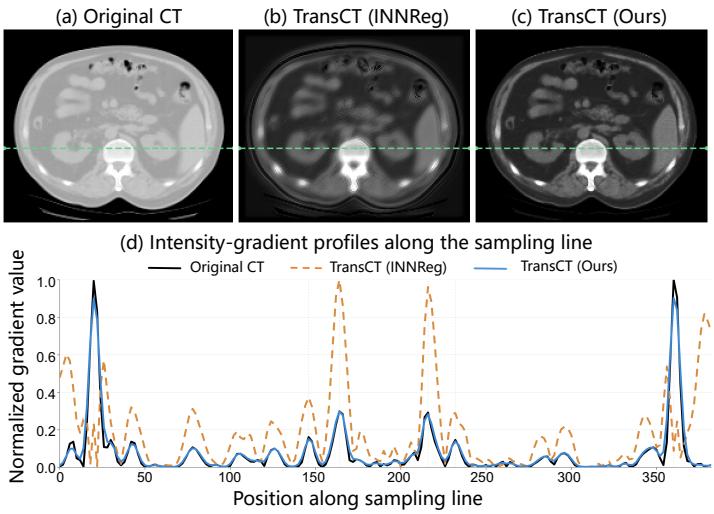

Figure 4: Comparison of transformed CT images and normalized gradient profiles along the indicated sampling line.

Table 3: Ablation study on the Learn2Reg CT–MR dataset. The shared-weight encoder is retained in all configurations. "w/o" and "w/" denote without and with MR-histogram conditioning, respectively.

Configuration	DSC ↑	HD95 ↓	ASSD ↓
RDP Baseline	0.597	12.099	4.620
+ TransNet w/o MR-hist	0.649	12.063	4.577
+ TransNet w/ MR-hist (Ours)	**0.658**	**11.740**	**4.324**

Ablation Study. We evaluate the contributions of TransNet and MR-histogram conditioning while retaining the shared-weight encoder in all configurations. Weight sharing is treated as a fundamental architectural constraint because it encourages CT, TransCT, and MR images to be represented by a common feature extractor, promoting structural features shared across modalities while suppressing modality-specific contrast variations. As shown in Table 3, the RDP Baseline achieves a DSC of 0.597, an HD95 of 12.099, and an ASSD of 4.620. The contrast transformation branch without MR-histogram conditioning (TransNet w/o MR-hist) improves the corresponding metrics to 0.649, 12.063, and 4.577. The complete configuration, TransNet w/ MR-hist, achieves the best performance, with a DSC of 0.658, an HD95 of 11.740, and an ASSD of 4.324. These results suggest that histogram-conditioned contrast transformation facilitates the learning of modality-invariant features in a common latent space, thereby promoting more precise anatomical correspondence.

6 Conclusion

This study targeted the challenges inherent in multi-modal medical image registration, with a particular focus on MRI-CT alignment, which is rendered difficult by modality-specific intensity discrepancies and substantial nonlinear anatomical deformations. We proposed a novel registration framework for multi-modal image alignment that leverages contrast transformation for improved structural correspondence estimation. By constraining the network with the MR intensity histogram, the CT images were transformed into MR-like contrast while preserving the underlying anatomy. This design enabled the shared-weight encoder to focus on structural features rather than contrast-induced variability. Extensive experiments on two challenging datasets demonstrated that our approach achieved state-of-the-art registration performance and exhibited strong generalizability to enhance established registration architectures.

AI USE STATEMENT

Generative AI tools were used solely for language polishing and improving the clarity of the manuscript. All scientific content, methodology, experiments, analyses, and conclusions were developed and verified by the authors.

REPRODUCIBILITY STATEMENT

Section 3 describes the architecture, feature-mixing schedule, and training objectives. Section 4 reports the dataset splits, preprocessing, construction of fixed training pairs, evaluation metrics, hardware, and optimization settings. Tables 1–3 provide the reported baseline comparisons, backbone-integration results, and ablation configurations. We will release our code upon acceptance.

REFERENCES

Mor Avi-Aharon, Assaf Arbelle, and Tammy Riklin Raviv. Differentiable histogram loss functions for intensity-based image-to-image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11642–11653, 2023.

Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

Yuan Chang, Zheng Li, and Wenzheng Xu. CGNet: A Correlation-Guided Registration Network for Unsupervised Deformable Image Registration. *IEEE Transactions on Medical Imaging*, 44: 1468–1479, 2024.

Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis*, 82:102615, 2022.

Songyu Chen, Sijuan Huang, Liwei Deng, Jing Wang, and Xin Yang. TransReg: Multi-modal medical image registration via image translation guidance. *Expert Systems with Applications*, pp. 130777, 2025.

Zeyuan Chen, Yuanjie Zheng, and James C Gee. TransMatch: a transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration. *IEEE transactions on medical imaging*, 43(1):15–27, 2023.

Kyobin Choo, Hyunkyung Han, Jinyeong Kim, Chanyong Yoon, and Seong Jae Hwang. Mono-Modalizing Extremely Heterogeneous Multi-modal Medical Image Registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 433–443. Springer, 2025.

Xueming Fu, Fenghe Tang, Rongsheng Wang, Yingtai Li, Lixia Han, Jian Lu, Zihang Jiang, and S Kevin Zhou. Medgmae: Gaussian masked autoencoders for medical volumetric representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hastings Greer, Lin Tian, François-Xavier Vialard, Roland Kwitt, Raul San Jose Estepar, and Marc Niethammer. CARL: A Framework for Equivariant Image Registration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26014–26023, 2025.

Mengjie Guo. Unsupervised multi-modal medical image registration via invertible translation. In *European Conference on Computer Vision*, pp. 22–38. Springer, 2024.

Xudong Guo, Peiyu Chen, Haifeng Wang, Zhichao Yan, Qinfen Jiang, Rongjiang Wang, and Ji Bin. UTMorph: A Hybrid CNN-Transformer Network for Weakly-supervised Multimodal Image Registration in Biopsy Puncture. *Medical Image Analysis*, pp. 103938, 2026.

Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical image analysis*, 16(7):1423–1435, 2012.

Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712, 2022.

Yibo Hu, Qi Zhang, Ziqi Zhao, Xiaoyi Ding, Wentao Li, Jijin Yang, Jianqi Sun, and Lisa X Xu. ARDMR: Adaptive recursive inference and representation disentanglement for multimodal large deformation registration. *Medical Image Analysis*, pp. 103844, 2025.

Zhengyong Huang and Yao Sui. Dual-Stream Pyramidal Attention Networks for Medical Image Registration. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2025.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

Miao Kang, Xiaojun Hu, Weilin Huang, Matthew R. Scott, and Mauricio Reyes. Dual-stream pyramid registration network. *Medical image analysis*, 78:102379, 2022.

Daniel Kim, Abdullah Shazly, Mohammed A Al-Masni, Dong-Hyun Kim, and Kanghyun Ryu. Cross-Modality Image Registration Via Generating Aligned Image Using Reference-Augmented Framework. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Lingke Kong, Chenyu Lian, Detian Huang, Yanle Hu, Qichao Zhou, et al. Breaking the dilemma of medical image-to-image translation. *Advances in Neural Information Processing Systems*, 34: 1964–1978, 2021.

Xiyu Luo, Haodong Li, Xinxing Cheng, He Zhao, Yang Hu, Xuan Song, and Tianyang Zhang. TraceTrans: Translation and Spatial Tracing for Surgical Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7698–7706, 2026.

Jinxin Lv, Zhiwei Wang, Hongkuan Shi, Haobo Zhang, Sheng Wang, Yilang Wang, and Qiang Li. Joint progressive and coarse-to-fine registration of brain MRI via deformation field integration and non-rigid feature fusion. *IEEE Transactions on Medical Imaging*, 41(10):2788–2802, 2022.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2021.

Tai Ma, Xinru Dai, Suwei Zhang, and Ying Wen. Pivit: Large deformation image registration with pyramid-iterative vision transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 602–612. Springer, 2023.

Mingyuan Meng, Dagan Feng, Lei Bi, and Jinman Kim. Correlation-aware Coarse-to-fine MLPs for Deformable Medical Image Registration. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9645–9654, 2024.

Tony CW Mok, Zi Li, Yunhao Bai, Jianpeng Zhang, Wei Liu, Yan-Jie Zhou, Ke Yan, Dakai Jin, Yu Shi, Xiaoli Yin, et al. Modality-agnostic structural image representation learning for deformable multi-modality medical image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11215–11225, 2024.

Anna Reithmeir, Veronika Spieker, Vasiliki Sideri-Lampretsa, Daniel Rueckert, Julia A Schnabel, and Veronika A Zimmer. From model based to learned regularization in medical image registration: A comprehensive review. *Medical Image Analysis*, pp. 103854, 2025.

Xinrui Song, Hengtao Guo, Xuanang Xu, Hanqing Chao, Sheng Xu, Baris Turkbey, Bradford J Wood, Ge Wang, and Pingkun Yan. Cross-modal attention for MRI and ultrasound volume registration. In *International conference on medical image computing and computer-assisted intervention*, pp. 66–75. Springer, 2021.

Xinrui Song, Hanqing Chao, Xuanang Xu, Hengtao Guo, Sheng Xu, Baris Turkbey, Bradford J Wood, Thomas Sanford, Ge Wang, and Pingkun Yan. Cross-modal attention for multi-modal image registration. *Medical Image Analysis*, 82:102612, 2022.

Xinrui Song, Xuanang Xu, Jiajin Zhang, Diego Machado Reyes, and Pingkun Yan. Dino-reg: efficient multimodal image registration with distilled features. *IEEE transactions on medical imaging*, 44 (9):3809–3819, 2025.

Haiqiao Wang, Dong Ni, and Yi Wang. ModeT: Learning deformable image registration via motion decomposition transformer. In *International conference on medical image computing and computer-assisted intervention*, pp. 740–749. Springer, 2023.

Haiqiao Wang, Dong Ni, and Yi Wang. Recursive deformable pyramid network for unsupervised medical image registration. *IEEE Transactions on Medical Imaging*, 43(6):2229–2240, 2024.

Yinsong Wang, Xinzhe Luo, Siyi Du, and Chen Qin. Adaptive Conditional Contrast-Agnostic Deformable Image Registration with Uncertainty Estimation. *IEEE Transactions on Medical Imaging*, 2026.

Yinuo Wang, Tao Guo, Weimin Yuan, Shihao Shu, Cai Meng, and Xiangzhi Bai. Mamba-based deformable medical image registration with an annotated brain MR-CT dataset. *Computerized Medical Imaging and Graphics*, 123:102566, 2025.

Shuwen Wei, Samuel W Remedios, Zhangxing Bian, Shimeng Wang, Junyu Chen, Yihao Liu, Bruno Jedynak, Tin YA Liu, Shiv Saidha, Peter A Calabresi, et al. Unsupervised OCT image interpolation using deformable registration and generative models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 661–671. Springer, 2025.

Peng Yuan, Jianmin Dong, Wei Zhao, Fei Lyu, Cheng Xue, Yudong Zhang, Chunfeng Yang, Zhan Wu, Zhiqiang Gao, Tianling Lyu, et al. LADDA: Latent Diffusion-based Domain-adaptive Feature Disentangling for Unsupervised Multi-modal Medical Image Registration. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Shenglong Zhou, Zhiwei Xiong, and Feng Wu. Test-time adaptation via style and structure guidance for histological image registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7677–7685, 2024.

Jingke Zhu, Boyun Zheng, Bing Xiong, Yuxin Zhang, Ming Cui, Deyu Sun, Jing Cai, Yaoqin Xie, and Wenjian Qin. SynMSE: A multimodal similarity evaluator for complex distribution discrepancy in unsupervised deformable multimodal medical image registration. *Medical Image Analysis*, 103: 103620, 2025.

Muhammad Zubair, Muzammil Hussain, Mousa Ahmad Albashrawi, Malika Bendechache, and Muhammad Owais. A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis. *Comput. Methods Programs Biomed.*, 272:109014, 2025.