# Learning Hierarchical Anatomy-Grounded Representation for Cross-Modal Registration

Jia Mi[1], Caiwen Jiang[1], Haoyue Yuan[1], Fan Li[1], Xiaoye Li[1], Kaicong Sun[1], and Dinggang Shen[1,2,3](✉)

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
[2] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
[3] Shanghai Clinical Research and Trial Center, Shanghai, China
Dinggang.Shen@gmail.com

**Abstract.** Accurate cross-modal deformable medical image registration is essential for integrating complementary clinical information across modalities such as MRI and CT. However, the significant modality gap makes it difficult to establish reliable correspondences from raw image appearances. Contrastive learning has emerged as a promising approach for extracting modality-agnostic features for correspondence estimation. However, existing methods typically treat voxels as independent instances in contrastive learning, without explicitly modeling hierarchical anatomical context and topology, which are critical for disambiguating correspondences and preserving deformation plausibility. To address this limitation, we propose the Hierarchical Anatomy-Grounded Representation (HAGR) framework. HAGR augments contrastive learning with cross-view feature reconstruction to encourage structurally consistent, anatomy-aware representations beyond isolated local features. It further learns a hierarchical feature space in a coarse-to-fine manner, where a matchability-aware coarse stage provides global anatomical context priors and a coarse-guided fine stage improves local discriminability through implicit hard-sample mining. Extensive experiments on cross-modal registration benchmarks demonstrate that HAGR achieves accurate and robust registration compared with existing registration methods, providing a generalizable and reusable foundation for reliable cross-modal information fusion. (https://github.com/jiaem1/HAGR)

**Keywords:** Cross-modal Registration · Representation Learning · Deep Learning.

## 1 Introduction

Different medical image modalities, such as MRI and CT, provide complementary tissue information for diagnosis, treatment planning, and image-guided interventions [3,6]. Their integration typically requires precise spatial alignment of anatomical structures across modalities. However, multi-modal scans are often acquired under heterogeneous conditions, including different scanners, time

points, and physiological states, introducing substantial anatomical misalignment. Accurate cross-modal deformable registration is therefore essential for reliable cross-modal information fusion.

Cross-modal registration critically depends on measuring anatomical alignment despite large cross-modality appearance gaps. Early approaches relied on hand-crafted similarity measures, such as mutual information [19] and MIND [7], but their limited anatomical awareness reduces robustness under complex deformations. Recent deep learning methods typically either directly regress deformation fields [16,12], or perform cross-modal image synthesis to bridge the appearance gap before registration [3,6]. While effective in specific scenarios, direct regression can overfit to particular modality/contrast combinations, and synthesis may introduce artifacts that compromise anatomical fidelity. Consequently, shifting the similarity measure from the original image space to a learned modality-invariant feature space has emerged as a promising alternative for generalizable cross-modal registration [2,13,5].

Most representation learning methods adopt contrastive objectives to enforce modality invariance. This constraint is typically imposed on local point-wise pairs by pulling features at corresponding cross-modal locations together and pushing non-corresponding features apart. However, anatomy exhibits hierarchical structured dependencies among localized points, ranging from global anatomical context and topology to local textures and boundary details [23,21]. As a result, purely point-wise contrastive learning does not explicitly enforce how contextual information should be organized or used for correspondence reasoning, which may limit its ability to disambiguate anatomically similar locations and support plausible deformation fields [13,9]. This motivates a stronger anatomy-grounded proxy task, where representations are learned through structured correspondence reasoning that requires the use of anatomical and topological context. Moreover, a coarse-to-fine formulation provides a natural way to encode these hierarchical dependencies, allowing macroscopic anatomical context to constrain and guide the interpretation of local geometric details.

In this paper, we propose a **H**ierarchical **A**natomy-**G**rounded **R**epresentation (HAGR) framework for cross-modal deformable registration. HAGR augments contrastive learning with cross-view feature reconstruction, using similarity-based reconstruction from cross-view candidates as a structured proxy task to learn registration-oriented anatomical representations. This reconstruction objective is further embedded into a hierarchical coarse-to-fine architecture to capture cross-scale context for jointly resolving large displacements and subtle distortions. At the coarse level, HAGR performs matchability-aware reconstruction to learn a global context prior, improving robustness to large-scale displacement. At the fine level, coarse-guided implicit hard-sample mining focuses learning on locally difficult regions, enhancing discriminability and sensitivity to subtle misalignments. We then perform hierarchical feature-space registration using these representations. Extensive experiments on multi-modal registration benchmarks show that HAGR achieves more accurate and robust alignment than both end-to-end multi-modal registration methods and prior representation-learning

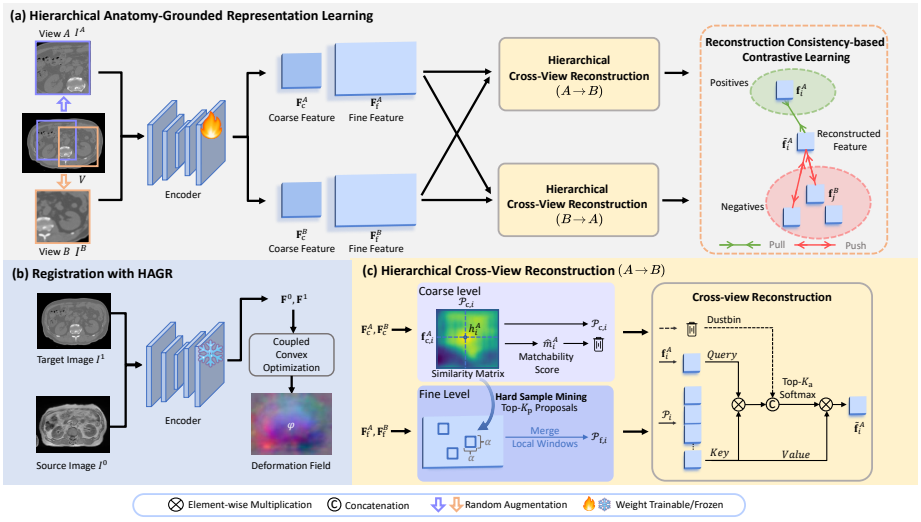


**Fig. 1.** Overview of the proposed framework: (a) hierarchical anatomy-grounded representation learning, (b) registration using learned HAGR features, and (c) hierarchical cross-view reconstruction.

baselines, suggesting that hierarchical anatomy-grounded representation learning provides a reusable foundation for reliable multi-modal deformable registration.

## 2 Method

As illustrated in Fig. 1, HAGR learns coarse-to-fine representations for registration. During training, given views $I^A$ and $I^B$, an encoder extracts coarse features $\mathbf{F}_{\mathrm{c}}^A, \mathbf{F}_{\mathrm{c}}^B \in \mathbb{R}^{C_1 \times D_1 \times H_1 \times W_1}$ at $\frac{1}{16}$ spatial resolution, and fine features $\mathbf{F}_{\mathrm{f}}^A, \mathbf{F}_{\mathrm{f}}^B \in \mathbb{R}^{C_2 \times D_2 \times H_2 \times W_2}$ at $\frac{1}{2}$ of the input resolution. These representations are jointly optimized via hierarchical cross-view reconstruction. At inference, registration is performed by optimizing a deformation field within this learned feature space. To avoid relying on paired multi-modal scans, we sample random 3D patches from a single CT volume $V$ and apply independent intensity and affine augmentations to generate views $I^A$ and $I^B$ to simulate heterogeneous cross-modal appearance variations [13,2,22].

### 2.1 Hierarchical Coarse-to-Fine Architecture

To couple global anatomical context with local geometric detail, our hierarchical architecture adapts the cross-view candidate pool $\mathcal{P}_{\ell,i}$ at each feature level $\ell \in \{\mathrm{c}, \mathrm{f}\}$. Let $\mathbf{f}_{\ell,i}^A$ denote the anchor feature at spatial location $i$ in view $A$, and define $\mathcal{P}_{\ell,i}$ as the cross-view candidate pool from view $B$ used to reconstruct $\mathbf{f}_{\ell,i}^A$. The same operations are applied symmetrically for $B \to A$ by exchanging the roles of $A$ and $B$.

**Coarse stage: matchability-aware global context modeling.** To learn representations that capture global anatomical structure and context for large displacement alignment, we augment cross-view reconstruction with matchability-aware modeling, which encourages the learned features to use global structural cues for distinguishing valid from invalid correspondences. Concretely, for each coarse-level anchor $\mathbf{f}^A_{\mathrm{c},i}$, the candidate pool is defined as the entire coarse feature set of the other view, i.e., $\mathcal{P}_{\mathrm{c},i} = \{\mathbf{f}^B_{\mathrm{c},j}\}$. We compute the strongest cross-view similarity response $h^A_i = \max_j \langle \mathbf{f}^A_{\mathrm{c},i}, \mathbf{f}^B_{\mathrm{c},j} \rangle$, and predict a matchability score $\hat{m}^A_i = \Psi_{\mathrm{mat}}\big([\mathbf{f}^A_{\mathrm{c},i}; h^A_i]\big) \in [0,1]$, where $\Psi_{\mathrm{mat}}$ is a lightweight predictor. This predicted matchability is used to construct a dustbin in cross-view reconstruction, introducing a soft rejection mechanism (Sec. 2.2) to discard features with weak or implausible correspondence.

**Fine stage: coarse-guided local discriminative refinement.** While the coarse stage captures global anatomical context, fine features preserve rich local morphology needed for precise dense correspondence. We leverage coarse correspondence cues to implement a hard-sample mining strategy for cross-view reconstruction, coupling two-level features to learn robust and discriminative representations. Specifically, for each fine-level anchor $\mathbf{f}^A_{\mathrm{f},i}$, we select the top-$K_p$ locations in view $B$ with the highest coarse feature similarity as coarse correspondence proposals. These proposals are then mapped to the fine grid, and the fine-level candidate pool $\mathcal{P}_{\mathrm{f},i}$ is defined as the set of fine-level feature vectors collected from local windows of size $\alpha \times \alpha \times \alpha$ centered at these proposal locations (Fig. 1(c)). By resolving reconstruction ambiguity among plausible but hard-to-distinguish local hypotheses, the fine representation is encouraged to encode more discriminative local anatomical cues, improving sensitivity to subtle misalignments.

### 2.2 Anatomy-Grounded Cross-View Reconstruction Learning

**Reconstruction with cross-view attention.** Given the hierarchical framework defined in Sec. 2.1, HAGR learns anatomy-grounded representations via cross-view reconstruction consistency in feature space. For a feature level $\ell \in \{\mathrm{c}, \mathrm{f}\}$, we reconstruct anchor feature $\mathbf{f}^A_{\ell,i}$ from view $A$ using its corresponding candidate pool $\mathcal{P}_{\ell,i}$ from view $B$ by a non-parametric attention mechanism:

$$\tilde{\mathbf{f}}^A_{\ell,i} = \sum_{\mathbf{u} \in \mathcal{P}_{\ell,i}} \alpha_{\ell,i}(\mathbf{u})\, \mathbf{u}, \quad \alpha_{\ell,i}(\mathbf{u}) \propto \exp\big(\langle \mathbf{f}^A_{\ell,i}, \mathbf{u} \rangle / \tau_\ell\big), \tag{1}$$

where $\tau_\ell$ is a temperature parameter. Without learnable projections, the reconstruction directly encourages the feature space to encode shared anatomy. In practice, we use sparse top-$K_a$ attention (i.e., softmax over the top-$K_a$ similarities within $\mathcal{P}_{\ell,i}$ for each anchor) to enforce correspondence sparsity and reduce feature over-smoothing. We augment the coarse attention logits with a dustbin by appending an unmatchability score, defined as $\log(1 - \hat{m}^A_i)$, to the cross-view similarity logits prior to the top-$K_a$ selection (dashed arrow in Fig. 1(c)). By

softly rejecting unmatchable anchors, this mechanism ensures that the reconstruction is primarily driven by valid candidate features.

**Objective function.** We optimize HAGR with contrastive objectives defined on reconstruction consistency at both feature levels. For level $\ell \in \{\mathrm{c}, \mathrm{f}\}$, the directional loss for $A \to B$ is

$$\mathcal{L}_{\mathrm{rec}}^{\ell}(A \to B) = -\frac{1}{|\mathcal{Q}_{\ell}^{A}|} \sum_{i \in \mathcal{Q}_{\ell}^{A}} \log \frac{\sum_{\mathbf{u} \in \mathcal{G}_{\ell,i}^{A}} \exp\big(s(\tilde{\mathbf{f}}_{\ell,i}^{A}, \mathbf{u})/\tau_{\ell}\big)}{\sum_{\mathbf{w} \in \mathcal{G}_{\ell,i}^{A} \cup \mathcal{N}_{\ell,i}^{A}} \exp\big(s(\tilde{\mathbf{f}}_{\ell,i}^{A}, \mathbf{w})/\tau_{\ell}\big)}, \tag{2}$$

where $s(\cdot, \cdot)$ denotes dot product, and $\mathcal{Q}_{\ell}^{A}$ denotes the valid anchors at level $\ell$, and $\mathcal{G}_{\ell,i}^{A}$ and $\mathcal{N}_{\ell,i}^{A}$ denote the positive and hard-negative feature sets for anchor $i$, respectively. At both levels, hard negatives are sampled from the most similar feature locations whose geometric distance to the anchor exceeds the corresponding ignore threshold ($\delta_{\mathrm{c}}$ for coarse and $\delta_{\mathrm{f}}$ for fine). For the positive set, the coarse level uses only the anchor itself, $\mathcal{G}_{\mathrm{c},i}^{A} = \{\mathbf{f}_{\mathrm{c},i}^{A}\}$, whereas the fine level uses a local neighborhood, with $\mathcal{G}_{\mathrm{f},i}^{A}$ defined as features within a small radius $\theta_{\mathrm{f}}$ to preserve local morphological continuity. The coarse matchability scores are supervised by binary cross-entropy: $\mathbf{m} \in \{0, 1\}$: $\mathcal{L}_{\mathrm{mat}} = \sum_{X \in \{A,B\}} \mathcal{L}_{\mathrm{BCE}}(\hat{\mathbf{m}}^{X}, \mathbf{m}^{X})$. The binary labels $\mathbf{m}^{X}$ come from known spatial augmentations, where anchors mapped inside the valid overlap are matchable and the rest unmatchable. This term trains the dustbin branch to softly reject anchors without valid correspondences, reducing noisy reconstruction gradients from non-overlapping regions. The final objective is

$$\mathcal{L} = \lambda_{\mathrm{c}} \mathcal{L}_{\mathrm{rec}}^{\mathrm{c}} + \lambda_{\mathrm{f}} \mathcal{L}_{\mathrm{rec}}^{\mathrm{f}} + \lambda_{\mathrm{m}} \mathcal{L}_{\mathrm{mat}}, \tag{3}$$

where $\mathcal{L}_{\mathrm{rec}}^{c}$ and $\mathcal{L}_{\mathrm{rec}}^{f}$ are the coarse- and fine-level instances of Eq. (2), respectively, and $\mathcal{L}_{\mathrm{rec}}^{\ell} = (\mathcal{L}_{\mathrm{rec}}^{\ell}(A \to B) + \mathcal{L}_{\mathrm{rec}}^{\ell}(B \to A))/2$. $\lambda_{\mathrm{c}}$, $\lambda_{\mathrm{f}}$, and $\lambda_{\mathrm{m}}$ are weighting parameters.

### 2.3 Registration with Learned Features

We use the learned features to measure cross-modal similarity for deformable registration. Specifically, we upsample the coarse features to the spatial resolution of the fine features and concatenate them to obtain the final feature map $\mathbf{F} \in \mathbb{R}^{(C_1+C_2) \times D_2 \times H_2 \times W_2}$. Given two images $I^0$ and $I^1$, let $\mathbf{F}^0$ and $\mathbf{F}^1$ denote their corresponding concatenated features. We construct a local cost volume $\mathbf{C}$ in the learned feature space as

$$\mathbf{C}(\mathbf{x}, \mathbf{d}) = \left\langle \mathbf{f}_{\mathbf{x}}^{1}, \mathbf{f}_{\mathbf{x}+\mathbf{d}}^{0} \right\rangle, \tag{4}$$

where $\mathbf{x}$ is a voxel location, $\mathbf{d}$ is a displacement within the search range of radius $N$, and $\mathbf{f}$ denotes the sampled feature vector at the specified location. We then perform registration using a multi-resolution pyramid strategy to improve the capture range and robustness. At each pyramid level $r$, features are extracted from the warped moving image and the fixed image, and an incremental deformation update $\varphi_r$ is computed using the coupled convex optimization scheme in

ConvexAdam [15]. The updates are composed across scales $\varphi_r^{\uparrow} = \mathrm{Up}(\varphi_r \circ \varphi_{r-1}^{\uparrow})$, where $\mathrm{Up}(\cdot)$ denotes upsampling. This is followed by a final instance optimization in feature space for fine-grained alignment, using a learning rate of 3 for 50 iterations.

## 3 Experiments and Results

### 3.1 Dataset and Implementation

**Datasets.** We utilize five datasets to evaluate intra-domain performance and cross-domain generalizability. (1) **Training:** We use 50 unpaired CT scans from Learn2Reg [8], 500 multi-center abdomen CT scans from AMOS22 [11], and 1204 CT scans from TotalSegmentator [20]. (2) **Intra-domain Evaluation:** 8 paired CT–MR cases from Learn2Reg serve as the primary benchmark. (3) **Generalization Evaluation:** We use the ThoraxCBCT [10] dataset (11 subjects), where each subject includes one lung fan-beam CT and two cone-beam CT scans acquired during image-guided radiotherapy, and an in-house Whole-Body dataset (20 subjects), where each subject includes one CT and four T1-weighted water-fat separated MR scans. CT scans from ThoraxCBCT are windowed using a lung setting (level $-600$, width 1500), while other CT scans use an abdominal setting (level 40, width 400). MR scans are normalized by the 0.5%/99.5% intensity quantiles. All images are resampled to an isotropic spacing of 2 mm. Registration accuracy is measured by DSC (Dice similarity coefficient) and HD95 (95th-percentile Hausdorff distance) on CT–MR datasets, and by TRE (Target Registration Error) and TRE30 (TRE on the top 30% landmarks with the largest initial error) on ThoraxCBCT. We additionally report the percentage of folding voxels, $\%|J_\varphi| \leq 0$, to assess deformation plausibility.

**Implementation.** HAGR is implemented in PyTorch [14] and trained on an NVIDIA A100 GPU with batch size $\mathcal{B} = 2$. We use a 3D ResNet backbone with two 3D FPN heads for coarse/fine features [18]. We set $C_1 = C_2 = 128$, $\alpha = 5$, $K_a = \{8, 64\}$ (coarse/fine), and $K_p = 3$. Optimization uses AdamW with learning rate 0.02. We warm up the encoder for 10,000 iterations with a standard contrastive loss, and then optimize Eq. 3 with $\lambda_{\mathrm{c}} = \lambda_{\mathrm{f}} = \lambda_{\mathrm{m}} = 1$.

### 3.2 Intra-domain Registration Performance

We first evaluate the registration accuracy on the abdominal CT–MR dataset Learn2Reg. We compare HAGR against seven competing methods including traditional baseline SyN [1], deep-learning-based baselines TransMorph [4] and DGMIR [12], and representation-based methods ConvexAdam [15], UAE [2], DINO-Reg [17], and Anatomix [5]. TransMorph and DGMIR are trained on unpaired training data with MIND loss and a regularization term.

As shown in Table 1, HAGR outperforms all competitors with an average DSC of 74.56% and the lowest HD95. TransMorph [4] struggles with cross-modality intensity shifts, resulting in limited accuracy (DSC of 41.85%). While

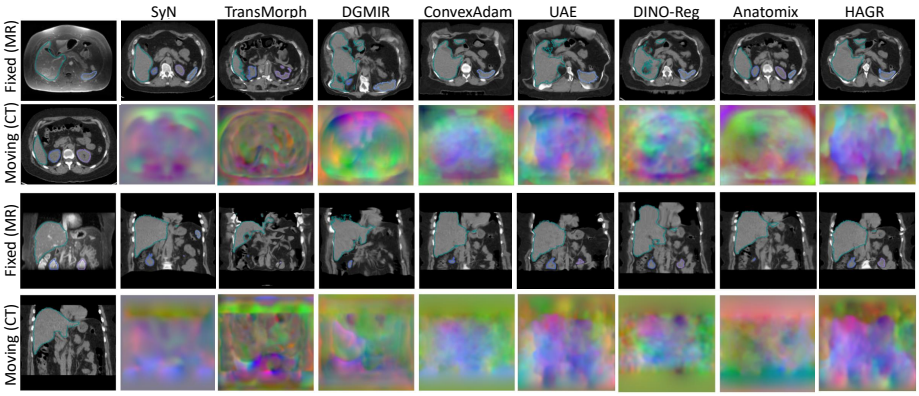


**Fig. 2.** Qualitative comparison on abdominal CT–MR registration in coronal and axial views. Segmentation contours are overlaid.

**Table 1.** Quantitative comparison of registration methods on abdominal CT–MR registration. Values are reported as mean ± standard deviation.

| **Method** | **DSC** (↑, %) | **HD95** (↓, mm) | $\%\lvert J_\varphi\rvert \leq 0$ | **Time** (s) |
|---|---|---|---|---|
| SyN [1] | $35.55 \pm 24.58$ | $19.24 \pm 14.29$ | $4.99 \pm 1.31$ | 17.3242 |
| TransMorph [4] | $41.85 \pm 26.33$ | $16.62 \pm 8.76$ | $36.87 \pm 2.59$ | 0.1684 |
| DGMIR [12] | $55.09 \pm 32.16$ | $11.89 \pm 10.32$ | $0.79 \pm 0.29$ | 1.4875 |
| ConvexAdam [15] | $66.58 \pm 28.28$ | $9.67 \pm 9.74$ | $4.53 \pm 2.26$ | 1.6306 |
| UAE [2] | $69.49 \pm 28.15$ | $7.77 \pm 9.12$ | $0.00 \pm 0.00$ | 1.7131 |
| DINO-Reg [17] | $71.23 \pm 25.05$ | $7.85 \pm 8.34$ | $5.13 \pm 3.02$ | 511.4721 |
| Anatomix [5] | $58.59 \pm 31.02$ | $12.21 \pm 11.67$ | $0.00 \pm 0.00$ | 3.3756 |
| HAGR | $74.56 \pm 27.50$ | $4.97 \pm 7.31$ | $0.00 \pm 0.00$ | 1.5680 |

DGMIR improves alignment by explicitly modeling modality gap, it still lags behind representation-based approaches. Among the representation-based methods, HAGR surpasses the runner-up (DINO-Reg) by 3.3% in DSC. These results indicate that anatomy-grounded features provide more robust alignment than generic self-supervised features. The qualitative comparisons in Fig. 2 show that competing methods often exhibit visible misalignment under large deformations and substantial modality gaps, whereas HAGR better preserves anatomical alignment.

### 3.3 Cross-domain Generalizability and Robustness

The learned modality-agnostic representation enables zero-shot transfer to unseen anatomies and modalities. We benchmark HAGR on the Whole-Body and ThoraxCBCT datasets against SyN and other representation-based methods that require no additional training or fine-tuning. Table 2 shows that HAGR

**Table 2.** Registration performance on two cross-domain datasets.

| | Whole-Body Dataset | | | ThoraxCBCT Dataset | | |
|---|---|---|---|---|---|---|
| Method | **DSC** (↑, %) | **HD95** (↓, mm) | $\%\lvert J_\varphi\rvert \leq 0$ | **TRE** (↓, mm) | **TRE30** (↓, mm) | $\%\lvert J_\varphi\rvert \leq 0$ |
| SyN | $13.98 \pm 6.32$ | $29.98 \pm 6.45$ | $4.11 \pm 0.74$ | $16.83 \pm 2.57$ | $21.78 \pm 3.14$ | $1.24 \pm 0.00$ |
| ConvexAdam | $58.59 \pm 28.74$ | $11.47 \pm 12.27$ | $4.25 \pm 1.72$ | $12.82 \pm 2.93$ | $20.03 \pm 2.82$ | $2.90 \pm 0.00$ |
| UAE | $56.08 \pm 16.34$ | $7.91 \pm 6.78$ | $0.00 \pm 0.00$ | $12.72 \pm 3.05$ | $23.67 \pm 4.12$ | $0.00 \pm 0.00$ |
| DINO-Reg | $57.16 \pm 16.05$ | $7.41 \pm 6.39$ | $1.99 \pm 0.83$ | $19.11 \pm 3.13$ | $24.28 \pm 3.22$ | $0.00 \pm 0.00$ |
| Anatomix | $52.87 \pm 31.43$ | $13.10 \pm 13.50$ | $0.00 \pm 0.00$ | $12.73 \pm 2.73$ | $23.43 \pm 4.16$ | $0.00 \pm 0.00$ |
| HAGR | $61.92 \pm 27.21$ | $9.18 \pm 11.02$ | $0.00 \pm 0.00$ | $12.34 \pm 2.58$ | $23.24 \pm 3.61$ | $0.00 \pm 0.00$ |

**Table 3.** Ablation study on abdominal CT–MR registration.

| Method Variant | **DSC** (↑, %) | **HD95** (↓, mm) | $\%\lvert J_\varphi\rvert \leq 0$ |
|---|---|---|---|
| Contrastive-Only | $64.31 \pm 29.61$ | $9.58 \pm 11.46$ | $0.00 \pm 0.00$ |
| Hard-Matching | $56.79 \pm 31.24$ | $10.29 \pm 10.91$ | $0.00 \pm 0.00$ |
| Coarse-Only | $39.78 \pm 29.64$ | $20.12 \pm 12.55$ | $0.00 \pm 0.00$ |
| Fine-Only | $64.58 \pm 32.08$ | $8.61 \pm 11.74$ | $0.00 \pm 0.00$ |
| HAGR | $74.56 \pm 27.50$ | $4.97 \pm 7.31$ | $0.00 \pm 0.00$ |

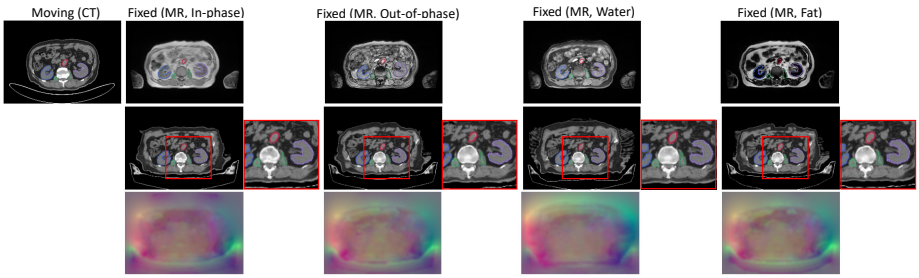


**Fig. 3.** Qualitative results on whole-body registration under varying MR contrasts.

provides strong zero-shot transfer performance, achieving the highest DSC on the Whole-Body dataset and the lowest mean TRE on ThoraxCBCT. On the Whole-Body dataset, HAGR achieves a 5.7% relative improvement in DSC compared to ConvexAdam. Qualitative results in Fig. 3 show that registering a single CT to four MR contrasts (Water, Fat, In-Phase, Out-of-Phase) yields highly consistent deformation fields, confirming that HAGR relies on anatomical geometry rather than intensity statistics.

### 3.4 Ablation Study

We analyze component contributions on the Learn2Reg dataset (Table 3). Replacing cross-view reconstruction with a standard contrastive loss leads to a clear performance drop, suggesting that point-wise feature contrast alone is insuffi-

cient for learning registration-oriented correspondence structure. Hard-Matching also performs worse than the proposed soft reconstruction, indicating that discrete correspondence assignments provide unstable supervision under ambiguous cross-modal matching. Removing either the coarse or fine branch further degrades performance, highlighting the importance of integrating coarse anatomical guidance with fine-level discriminability for robust deformable registration.

## 4 Conclusion

In this paper, we proposed HAGR, a hierarchical anatomy-grounded representation framework for cross-modal deformable registration. HAGR augments contrastive learning with cross-view reconstruction consistency within a progressive coarse-to-fine architecture. By training with a structured cross-view reconstruction task, HAGR shapes modality-agnostic representations toward registration-relevant correspondence reasoning, coupling coarse anatomical context with fine local discriminability. Experiments demonstrate that HAGR improves registration accuracy and deformation plausibility over competing methods, suggesting its potential as a reusable foundation for cross-modal deformable registration.

**Acknowledgments.** This work was supported in part by National Natural Science Foundation of China (grant numbers U23A20295, 62131015, 82441023, 82394432), the China Ministry of Science and Technology (STI2030-Major Projects-2022ZD0209000, STI2030-Major Projects-2022ZD0213100), The Key R&D Program of Guangdong Province, China (grant number 2023B0303040001), and HPC Platform of ShanghaiTech University.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Avants, B.B., Tustison, N., Song, G., et al.: Advanced normalization tools (ANTS). Insight j **2**(365), 1–35 (2009)
2. Bai, X., Bai, F., Huo, X., Ge, J., Lu, J., Ye, X., Yan, K., Xia, Y.: UAE: Universal anatomical embedding on multi-modality medical images. arXiv preprint arXiv:2311.15111 (2023)
3. Cao, X., Yang, J., Gao, Y., Guo, Y., Wu, G., Shen, D.: Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis. Medical image analysis **41**, 18–31 (2017)
4. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: TransMorph: Transformer for unsupervised medical image registration. Medical image analysis **82**, 102615 (2022)
5. Dey, N., Billot, B., Wong, H.E., Wang, C.J., Ren, M., Grant, P.E., Dalca, A.V., Golland, P.: Learning general-purpose biomedical volume representations using randomized synthesis. arXiv preprint arXiv:2411.02372 (2024)

6. Han, R., Jones, C.K., Lee, J., Wu, P., Vagdargi, P., Uneri, A., Helm, P.A., Luciano, M., Anderson, W.S., Siewerdsen, J.H.: Deformable MR-CT image registration using an unsupervised, dual-channel network for neurosurgical guidance. Medical image analysis **75**, 102292 (2022)
7. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A.: MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical image analysis **16**(7), 1423–1435 (2012)
8. Hering, A., Hansen, L., Mok, T.C., Chung, A.C., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., 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)
9. Huang, W., Yang, H., Liu, X., Li, C., Zhang, I., Wang, R., Zheng, H., Wang, S.: A coarse-to-fine deformable transformation framework for unsupervised multi-contrast mr image registration with dual consistency constraint. IEEE Transactions on Medical Imaging **40**(10), 2589–2599 (2021)
10. Hugo, G.D., Weiss, E., Sleeman, W.C., Balik, S., Keall, P.J., Lu, J., Williamson, J.F.: A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer. Medical physics **44**(2), 762–771 (2017)
11. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in neural information processing systems **35**, 36722–36732 (2022)
12. Le, G., Shu, Y., Qiao, L., Yang, L., Xiao, B., Li, W., Gao, X.: DGMIR: Dual-guided multimodal medical image registration based on multi-view augmentation and on-site modality removal. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 152–162. Springer (2025)
13. Mok, T.C., Li, Z., Bai, Y., Zhang, J., Liu, W., Zhou, Y.J., Yan, K., Jin, D., Shi, Y., Yin, X., 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)
14. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
15. Siebert, H., Großbröhmer, C., Hansen, L., Heinrich, M.P.: ConvexAdam: Self-configuring dual-optimisation-based 3d multitask medical image registration. IEEE Transactions on Medical Imaging (2024)
16. Song, X., Chao, H., Xu, X., Guo, H., Xu, S., Turkbey, B., Wood, B.J., Sanford, T., Wang, G., Yan, P.: Cross-modal attention for multi-modal image registration. Medical Image Analysis **82**, 102612 (2022)
17. Song, X., Xu, X., Yan, P.: Dino-Reg: General purpose image encoder for training-free multi-modal deformable medical image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 608–617. Springer (2024)
18. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8922–8931 (2021)
19. Viola, P., Wells III, W.M.: Alignment by maximization of mutual information. International journal of computer vision **24**(2), 137–154 (1997)

20. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: TotalSegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
21. Wu, L., Zhuang, J., Chen, H.: VoCo: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22873–22882 (2024)
22. Yan, K., Cai, J., Jin, D., Miao, S., Guo, D., Harrison, A.P., Tang, Y., Xiao, J., Lu, J., Lu, L.: SAM: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. IEEE Transactions on Medical Imaging **41**(10), 2658–2669 (2022)
23. Zhuang, J., Wu, L., Wang, Q., Fei, P., Vardhanabhuti, V., Luo, L., Chen, H.: MiM: Mask in mask self-supervised pre-training for 3d medical image analysis. IEEE Transactions on Medical Imaging (2025)