KGT-Reg: Knowledge-Guided Multimodal Framework for Medical Image Registration

Yuhe Dai[1,2,3], Zhiyong Huang[1,2], Xiaohong Wang[3], and Weimin Huang[3*]

[1]School of Computing, National University of Singapore, Singapore
[2] National University of Singapore Chongqing Research Institute, China
[3]Institute of Advanced Intelligence and Computing (IAIC), Agency for Science, Technology and Research(A*Star), Singapore
*wmhuang@a-star.edu.sg

Abstract. Deformable medical image registration (DIR) benefits from both pixel-level correspondence modelling and high-level semantic understanding of anatomical structures. However, most learning-based methods remain purely data-driven and fail to leverage rich domain knowledge describing anatomical structures and inter-organ relations. In this paper, we present KGT-Reg, a unified multimodal[1] framework that integrates knowledge graph (**KG**) embedding and **t**extual semantics into deep **reg**istration networks. KGT-Reg fuses visual features extracted from paired medical volumes with semantic embeddings derived from organ-level textual descriptions and graph neural encodings of anatomical KGs. Specifically, the KG encodes global relation priors across anatomical entities to represent organ adjacency and spatial topology, while text embeddings capture localized semantic context. A graph-enhanced fusion module further dynamically modulates feature decoding to promote spatially consistent deformations. Experiments on four benchmark datasets — OASIS, IXI, LPBA40 (brain MRI), and Learn2Reg 2020 (abdomen CT) — demonstrate consistent improvements over state-of-the-art baselines, validating the effectiveness of structured knowledge integration and text aggregation for multimodal DIR.

Keywords: Medical Image Registration, Knowledge Graph, Text Embedding.

1 Introduction

Medical image registration is a foundational problem in computational anatomy and image-guided intervention, aiming to establish dense spatial correspondence between scans across subjects, time points, or modalities [1], which underpins applications such as longitudinal disease tracking, surgical navigation, and multimodal diagnosis. Classical methods (e.g., SyN [2], Demons [3]) optimize similarity metrics under regularization constraints but suffer from high computational costs and sensitivity to initialization. Deep learning-based networks address this by predicting deformation fields directly from image pairs [4], with VoxelMorph [5] demonstrating that convolutional encoder–decoder architectures can achieve smooth, diffeomorphic registration with orders-of-magnitude speedup. Subsequent works have further incorporated probabilistic

[1] Here, "multimodal" refers to multiple representation modalities (image, text, and graph), not multiple imaging modalities (e.g., T1/T2 MRI). * Corresponding Author.

regularization, multi-scale pyramids, and transformer-based representations [6–9]. Despite this progress, existing deep registration methods rely primarily on voxel intensity patterns and local gradients optimized through pixel-wise similarity losses, limiting generalization to large anatomical variability or pathological deformations. More critically, most deep registration networks lack anatomical awareness: they align intensities without explicit knowledge of underlying structure, producing spatially smooth yet anatomically implausible transformations—such as organ overlap or topology violations—particularly problematic in abdominal registration [10].

To inject semantic knowledge, vision–language models (VLMs) [11], e.g., MedCLIP [12], GLORIA [13], MedVLM [14], have been explored in medical imaging, but text embeddings alone lack the relational structure needed to encode inter-organ spatial relationships and topological constraints. Knowledge graphs (KGs) [15], reasoned over via graph neural networks (GNNs) [16] through message passing, offer a complementary solution by formalizing relational priors over anatomical entities, yielding improved interpretability and generalization. While KGs have proven effective in segmentation, diagnosis, and report generation [17–20], their application to DIR remains unexplored, despite its fundamental reliance on anatomical spatial relationships.

Thus, we propose KGT-Reg, a unified multimodal framework that integrates image, text, and knowledge-graph pathways to achieve semantically consistent and correspondence-disambiguated registration. KGT-Reg introduces two key mechanisms to bridge data-driven registration with structured anatomical knowledge: (1) text-guided semantic alignment, which encodes anatomical structure descriptions into dense embeddings to enable structure-aware, clinically prioritized alignment; and (2) a KG relational learning module, which propagates dataset-level organ-wise spatial priors to condition the deformation decoder, producing displacement fields that respect anatomical topology. The state-of-the-art results on OASIS [21], IXI [22], LPBA40 [23], and Learn2Reg 2020 [10] validate our method's effectiveness. Our main contributions are:

- We propose a novel unified multimodal registration framework integrating knowledge from both textual semantics and structured KG into deep DIR, which is the first attempt to use KG relational priors to condition deformation-field decoding.
- We propose a text-prompted semantic guidance module that encodes anatomical entities into semantic embeddings, enriching visual features with high-level structural context for structure-aware registration.
- We propose a KG-based anatomical encoding module capturing inter-organ spatial relationships and hierarchical dependencies, enabling registration to incorporate population-level anatomical topology constraints.
- We generate state-of-the-art results on four publicly available datasets, demonstrating the effectiveness of KGT-Reg for DIR compared to existing methods.

2 Method

2.1 Network Overview

Given a moving image I_m and fixed image I_f, KGT-Reg predicts a dense deformation field $\phi(x) = x + u(x)$ over domain $\Omega \subset \mathbb{R}^{H\times W\times D}$, warping I_m into alignment with I_f

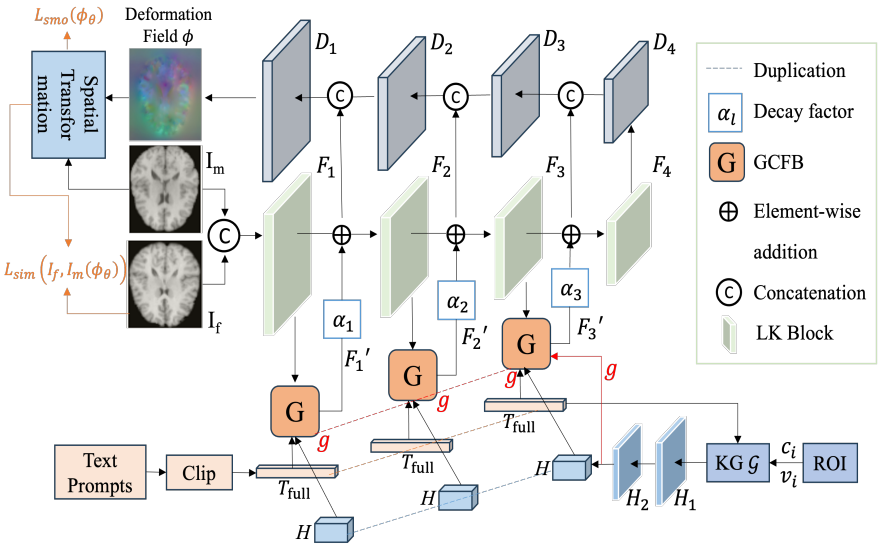

Fig. 1. Illustration of the proposed KGT-Reg model, where blocks in orange and blue distinguish text and KG branches. T_{full} and H are the full text- and KG-embedding matrices, and g is the global graph vector (Sec. 2.2). Dashed arrows indicate that T_{full}, H, and g are computed once and broadcast to every GCFB module, while c_i, v_i are node features used during KG construction.

to yield the deformed output $I_m \circ \phi$. The framework follows an encoder–decoder architecture augmented with graph-conditioned multimodal fusion (Fig. 1), comprising three synergistic components: (1) a large-kernel (LK) [24] convolutional encoder–decoder backbone for multi-scale spatial feature extraction; (2) a text encoder producing semantic embeddings for anatomical structures; and (3) a GNN-based KG encoder generating relational embeddings of organ-wise dependencies. These three modalities are fused via a Graph-Conditioned Fusion Block (GCFB) using feature-wise linear modulation (FiLM) [25], allowing semantic and relational cues to dynamically modulate feature extraction. Moreover, the encoder comprises four LK blocks with progressive downsampling and doubling channel width, while the decoder mirroring this with skip connections propagating both spatial detail and multimodally conditioned features. As shown in Fig. 2, multiscale fusion through stacked GCFBs ensures hierarchical integration of semantic and relational priors. At each level l, the fused feature map is modulated as $F_l^* = \alpha_l \cdot F_l'$, where F_l' is the feature map after GCFB, and the adaptive coefficient α_l controls the relative influence of multimodal conditioning, it increases with depth to reinforce high-level semantic reasoning while preserving low-level spatial precision. A convolutional regression head then predicts ϕ, applied via a spatial transformer to produce the registered image. The network is trained by minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{sim}}\left(I_f, I_m \circ \phi\right) + \mathcal{L}_{\text{aux}}\left(J_f, J_m \circ \phi\right) + \lambda \mathcal{L}_{\text{reg}}(\phi), \tag{1}$$

where the scalar coefficient λ balances accuracy against regularity. Here, we employ the normalized cross-correlation (NCC) or mean square error (MSE) (data-dependent)

as the similarity metric. We use a Dice loss on segmentation overlap to promote anatomical consistency and an L2 norm of the spatial gradients to enforce smoothness.

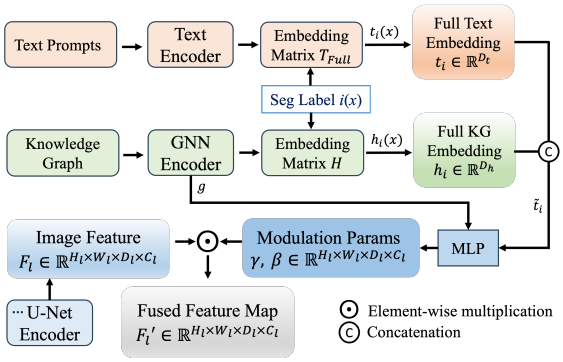

Fig. 2. The proposed GCFB module. Label *i(x)* retrieves graph/text embedding from matrix *H/T*.

2.2 Multimodal Encoding

Backbone for Image Encoding We adopt LKU-Net [24] as our backbone for its ability to capture both fine-grained and large-scale anatomical deformations via an enlarged effective receptive field at low computational cost. Formally, each LK block processes the input feature map F through four parallel convolutional paths with distinct kernel sizes and aggregates them via elementwise summation:

$$F_{\text{out}} = \psi(\mathcal{F}\text{conv}_{k\times k\times k}(F) + \mathcal{F}\text{conv}_{3\times 3\times 3}(F) + \mathcal{F}\text{conv}_{1\times 1\times 1}(F) + F), \quad (2)$$

where k is the LK size (set to 5), and $\psi(\cdot)$ shows a nonlinear activation (e.g., ReLU).

Textual Semantic Encoding We use a frozen CLIP text encoder [26] to generate semantic embeddings for anatomical structures via descriptive natural-language prompts of the form: "Magnetic Resonance Imaging (MRI) of the [i] region of the brain." Each structure i yields a textual vector $t_i \in \mathbb{R}^{D_t}$, with D_t denotes the embedding dimension; embeddings for M structures are aggregated as $T_{\text{full}} = [t_0, t_1, \dots, t_M]^T \in \mathbb{R}^{(M+1)\times D_t}$, where each row refers to one anatomical region, and t_0 is a distinct background embedding. Unlike one-hot encodings, CLIP embeddings capture rich inter-structural semantic relationships, providing expressive anatomical context for downstream fusion.

Anatomical KG Construction and Encoding We construct a dataset-level static KG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from training-set with segmentation masks, enabling the model to reason over global anatomical context rather than relying solely on local image-text features. Each node $v_i \in \mathcal{V}$ represents an anatomical structure with feature vector $x_i = [t_i, \bar{v}_i, \bar{c}_i]$, where $\bar{v}_i$ and $\bar{c}_i$ are the mean voxel volume and centroid. Computing dataset averages over the training split stabilizes the prior and reduces noise from individual segmentation errors. Moreover, the edge weights encode inter-organ spatial proximity:

$$A_{ij} = \exp(-d_{ij}/\sigma), d_{ij} = \| \bar{c}_i - \bar{c}_j \|, \quad (3)$$

with σ controls the sensitivity to distance and $A_{ii} = 1$, yielding a static adjacency matrix $A \in \mathbb{R}^{M\times M}$ capturing average spatial topology across subjects. The graph is then encoded by a two-layer Graph Attention Network (GAT) [27] for propagation:

$$H^{(1)} = \mathrm{GATConv}(X, A),\ H^{(2)} = \mathrm{GATConv}(H^{(1)}, A), \tag{4}$$

producing node embeddings $H = [h_1, \ldots, h_M]$. A global graph embedding $g = \frac{1}{M}\sum_i h_i$ summarizes global anatomical context and serves as a conditioning input to the GCFB.

2.3 Graph-conditioned Text-image Fusion

The GCFB integrates visual, textual, and relational features at multiple levels (Fig. 2). A mapping function is implemented at the voxel level through indexing. Given segmentation label $i(x)$, voxel-wise graph and text embeddings are retrieved as $h_i(x) = H[i(x)]$ and $t_i(x) = T_{\text{full}}[m(x)]$, then aggregated into a unified conditioning vector $\tilde{t}_i(x) = \mathrm{Concat}(t_i(x), h_i(x)) \in \mathbb{R}^{D_t+D_h}$, encoding both semantic identity ("what the structure is") and relational context ("how it relates to others"). Then, we employ a FiLM mechanism to dynamically adjust the features, and a multi-layer perceptron (MLP) is used to transform this vector—augmented with global constraint vector g for topological consistency—into modulation parameters:

$$(\gamma_l, \beta_l) = \mathrm{MLP}([\tilde{t}_i, g]), \tag{5}$$

where $\gamma_l, \beta_l \in \mathbb{R}^{C_l}$ are scale and shift coefficients, C_l is the channel dimension at level l of the U-Net. Concatenating g with local conditioning vectors ensures global topological coherence, preventing local deformations from violating structural continuity across anatomical levels. They modulate the image feature map via:

$$F_l' = \gamma_l \odot F_l + \beta_l, \tag{6}$$

where F_l represents the image feature map at level l, F_l' is the fused feature map, and $\odot$ denotes elementwise multiplication. In summary, GCFB extends voxel-level conditioning by combining the organ-specific embeddings from the text encoder and KG.

3 Experiments and Results

3.1 Setup

Datasets We evaluated KGT-Reg on four public datasets: three brain MRI datasets and an abdominal CT one. Segmentation labels were paired with anatomical names as text prompts; missing labels were generated via FreeSurfer [28] and the Neurite package [29]. For brain MRI, we used OASIS (Learn2Reg 2021 Task 3 [30]), comprising 414 T1-weighted volumes with 35 structure labels resampled to 160×192×224 and split into 394 training scans and 19 validation and ranking pairs; IXI [22], containing 576 T1-weighted scans with 38 structures cropped to 160×192×224 and split 403/58/115 for train/val/test; and LPBA40 [23], with 40 scans covering 56 labeled regions resampled

Table 1. Registration performance comparison on the brain and abdominal datasets.

Method	OASIS			LPBA40			IXI			Abdomen CT (Affine Dice 0.31)		
	Dice%	HD95	$\lvert J\rvert_{<0}$%	Dice%	HD95	$\lvert J\rvert_{<0}$%	Dice%	HD95	$\lvert J\rvert_{<0}$%	Dice%	HD95	$\lvert J\rvert_{<0}$%
SyN [2]	78.0	1.60	-	66.5	6.25	-	63.9	6.39	-	/	/	/
NiftyReg [34]	78.5	1.57	0.10	66.9	6.33	0.14	64.0	4.86	0.08	/	/	/
VoxelMorph [5]	84.7	1.55	1.24	64.2	6.81	0.76	72.6	3.75	1.52	38.8	25.2	3.96
TransMorph [8]	86.2	1.43	1.56	63.7	6.56	0.67	74.6	3.67	1.57	40.1	23.8	5.87
Im2Grid [32]	89.4	1.24	0.01	71.0	6.19	0.01	76.2	3.36	0.01	52.8	20.6	2.53
Fourier-Net [31]	86.0	1.37	0.02	67.2	6.72	0.25	76.7	**2.89**	0.02	40.7	23.7	2.85
LapIRN [7]	86.1	1.51	0.01	**73.6**	6.12	0.02	76.5	3.21	0.34	53.5	20.5	3.09
LKU-Net [24]	88.6	1.25	0.11	68.7	6.48	0.23	75.7	3.02	0.14	50.9	20.1	3.67
SACB-Net [33]	/	/	/	73.1	**5.86**	0.02	76.9	3.13	0.08	58.8	**18.3**	3.93
TextSCF [36]	90.1	1.22	0.12	/	/	/	/	/	/	**60.8**	22.4	3.31
T-Reg	91.4	1.25	0.11	72.5	6.29	0.13	78.6	2.98	0.13	59.9	20.5	3.26
KGT-Reg	**91.7**	**1.20**	0.12	73.1	6.22	0.13	**79.2**	2.92	0.11	60.7	19.8	3.19

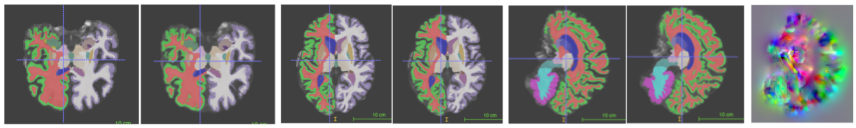

Fig. 3. One of the qualitative results on the OASIS dataset in blend display (before and after registration), in axial/coronal/sagittal dimension, with an example of the RGB displacement field (each channel encodes the displacement magnitude along one spatial axis.

to 160×192×160 and split 25/5/10. For abdominal CT, we used the Learn2Reg 2020 [30] Task 3 dataset [10], comprising 30 scans with 13 organ labels, where the Hounsfield Units was clamped to [-500, 800], normalized to [0, 1], resampled to 2 mm isotropic voxel spacing at 192×160×256, and split into 20/3/7 for train/val/test. This dataset evaluates generalization beyond neuroimaging under large organ-scale variation.

Implementation All experiments were implemented in PyTorch on four NVIDIA Quadro GV100 GPUs. The network was trained for 500 epochs using the Adam optimizer (batch size 1, learning rate 1×10^{-4}; training: ~18h, inference: ~0.9s/vol; memory: ~10GB). Regularization weight λ and initial feature channels C were tuned per dataset. Attenuation coefficients $\alpha_{1/2/3}$ were set empirically to 0.2, 0.4, and 0.6, gradually increasing multimodal influence from shallow to deep layers, preventing premature semantic injection that could disrupt low-level texture extraction.

Baseline and Evaluation Metrics We compared KGT-Reg against classical iterative algorithms, CNN- and Transformer-based methods, and hybrid models. Quantitative results were taken from official publications or leaderboards where available, with remaining baselines reproduced and fine-tuned using recommended configurations. Registration performance was evaluated using Dice Similarity Coefficient (DSC), 95 percentile Hausdorff Distance (HD95), and the percentage of negative values of the Jacobian determinant (% $\mid J_\phi \mid < 0$) to assess deformation field diffeomorphism quality.

3.2 Results

Brain MRI Registration Table 1 compares KGT-Reg against competing methods on three brain MRI datasets. On OASIS, KGT-Reg achieves the best Dice of 91.7%, lowest HD95 (1.20 mm), and reasonable folding ratio (0.12%), demonstrating that integrating semantic and relational priors improves alignment accuracy and encourages smooth and anatomically consistent deformations. Compared to the ablated T-Reg (without KG), the full model yields a further +0.3% Dice gain with reduced folding, confirming that KG-encoded relational topology provides meaningful structural regularization. On IXI, KGT-Reg again achieves the best Dice (79.2%) and second-best HD95 (2.92 mm). On LPBA40, KGT-Reg ranks second, slightly behind LapIRN, which benefits from an explicit multi-resolution similarity pyramid. Nonetheless, consistent performance across all datasets reflects the generalization of our multimodal fusion mechanism. Fig. 3 visualizes an OASIS case: the blended overlay (moving vs. fixed intensities before/after registration) across views, alongside the RGB displacement field. The reduced color bleeding and sharper anatomical boundaries in the post-registration blend confirm the smooth, locally varying deformation indicated by the low folding ratio in Table 1.

Abdomen CT Registration On the abdominal CT dataset (Table 1), featuring large inter-subject variability and complex organ boundaries, KGT-Reg still achieves second-best Dice and HD95, significantly outperforming classical and CNN-based methods. Compared to T-Reg, incorporating the KG yields consistent gains, demonstrating the stabilizing role of relational constraints under complex multi-organ deformations.

Comparison with LKU-Net and TextSCF Relative to the LKU-Net backbone, KGT-Reg achieves average Dice improvements of +3–4% across datasets with visibly smoother deformation maps, indicating that the model learns to respect semantic boundaries and organ adjacency beyond intensity fitting. Compared to TextSCF [36], another text-driven method, KGT-Reg integrates textual features at multiple hierarchical levels within the pathway, achieving a +1.6% Dice improvement on OASIS. While TextSCF demonstrates the value of textual guidance, KGT-Reg advances this paradigm through unified multiscale fusion of both textual and relational priors.

Ablation Study Table 2 investigates text encoder choice and prompt formulation on Abdomen CT. One-hot structure encoding yields a baseline Dice of 59.2%; substituting a CLIP encoder with a generic prompt slightly improves performance, and refining to a precise, modality-specific anatomical prompt achieves the best Dice of 60.5%, confirming that providing both anatomical and imaging modality context enhances structural understanding. Replacing with ChatGPT for the same modality-specific prompt leads to a performance drop, highlighting the importance of CLIP's image-text concept.

Meanwhile, Table 3 examines GCFB placement on OASIS: both encoder-only and decoder-only fusion variants underperform the full model but surpass LKU-Net and TextSCF, underscoring the benefit of KG-conditioned fusion throughout the network.

Table 4(a) further validates GCFB generalizability by applying our fusion module to DMR [35]: even with a weaker backbone, performance consistently improves with and w/o KG integration, confirming the modularity and effectiveness of our approach. Moreover, Table 4(b) explores the contribution of each component. Ablation on two

datasets, comparing with/without Dice configurations, disentangles architectural contributions from segmentation supervision. While adding Dice loss to LKU-Net yields gains, KGT-Reg w/o Dice still outperforms it, confirming that the improvements stem from the structured KG and text encoder rather than supervision alone. We also further add a KG-only ablation (without text prompt) –"KG-Reg" in Table 4(b). And KGT-Reg's consistent improvement confirms the unique value of both text and KG guidance.

Table 2. Ablation study on the impact of textual embeddings.

Text Prompt		Embedding	Dice(%)
i.	-	One-Hot	59.2
ii.	'Picture of [i]'	CLIP	59.9
iii.	'Picture of [i] in human abdomen.'	CLIP	60.2
iv.	'Computerized tomography of [i] in human abdomen'	CLIP	60.5
v.	'Computerized tomography of [i] in human abdomen'	ChatGPT	58.1

Table 3. Ablation study on the use of GCFB at different stages.

Model	Dice(%)	$\%\lvert J_\phi\rvert \leq 0$
Text fusion (without KG) at Encoder only	89.7	0.17±0.26
Text fusion (without KG) at Decoder only	89.2	0.14±0.06
GCFB at Encoder only	90.6	0.13±0.19
GCFB at Decoder only	90.0	0.13±0.08

Table 4. Ablation study on (a) a different backbone. (b) contribution of each component.

a. Model	OASIS		LPBA40	
	Dice(%)	$\%\lvert J_\phi\rvert \leq 0$	Dice(%)	$\%\lvert J_\phi\rvert \leq 0$
DMR [35]	79.3	1.02±0.44	67.5	0.62±0.33
T-DMR	80.8	0.98±0.43	68.9	0.56±0.33
KGT-DMR	81.6	0.94±0.38	69.0	0.45±0.29
b. Model	**OASIS**		**Abdomen CT**	
LKU-Net$_{\text{w/o Dice}}$	88.6	0.11±0.05	50.9	3.67±1.29
LKU-Net$_{\text{Dice}}$	89.2	0.10±0.07	55.2	3.43±1.00
KGT-Reg$_{\text{w/o Dice}}$	90.0	0.12±0.08	57.4	3.40±1.13
KGT-Reg$_{\text{Dice}}$	**91.7**	0.12±0.07	**60.7**	3.19±1.05
T-Reg$_{\text{Dice}}$	91.4	0.11±0.04	59.9	3.26±1.27
KG-Reg$_{\text{Dice}}$	91.4	0.12±0.10	60.1	3.23±1.12

4 Discussion and Conclusion

In our developing work, there are two primary limitations. First, although there are many open-source segmentation models and even some manually labelled ground truth exists, the KG quality somewhat depends on such segmentation accuracy, and errors in node statistics may bias the anatomical prior. Second, while static KG is lightweight and easy for use, such static adjacency matrix may not capture severe pathological rearrangements such as post-surgical resection. Thus, we are further investigating dynamic KG with per-case graph updates to mitigate this.

In this paper, we introduced KGT-Reg, a novel multimodal framework that integrates textual semantics and knowledge graph representations to bridge data-driven registration with anatomical reasoning. Evaluated on four benchmark datasets, KGT-Reg consistently outperforms most state-of-the-art baselines with strong spatial consistency and multi-organ generalization. Ablation studies confirm the role of each component: the KG provides global relational constraints, the text encoder supplies localized semantic context, and the hierarchical GCFB integrates both local appearance cues and global topological priors throughout the encoding–decoding pathway. Besides quantitative improvements, KGT-Reg represents a paradigm shift toward explainable design and knowledge-aware registration, which is particularly valuable in clinical applications.

Acknowledgments. This work was partly supported by HHP IAF-PP (grant number H25J6a0148).

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

References

1. D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Phys Med Biol*, vol. 46, no. 3, pp. R1-45, Mar. 2001.
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," *Med Image Anal*, vol. 12, no. 1, pp. 26–41, Feb. 2008.
3. T. Vercauteren *et al.*, "Symmetric log-domain diffeomorphic Registration: a demons-based approach," *Med Image Comput Comput Assist Interv*, vol. 11, no. Pt 1, pp. 754–761, 2008.
4. G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, no. 1, p. 8, Jan. 2020.
5. 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.
6. A. V. Dalca *et al.*, "Unsupervised Learning of Probabilistic Diffeomorphic Registration for Images and Surfaces," *Medical Image Analysis*, vol. 57, pp. 226–236, Oct. 2019.
7. T. C. W. Mok and A. C. S. Chung, "Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks," *MICCAI 2020,* Lima, Peru, October 4–8, 2020, Proceedings, Part III 23, pages 211–221. Springer, 2020.
8. J. Chen *et al.*, "TransMorph: Transformer for unsupervised medical image registration," *Medical Image Analysis*, vol. 82, p. 102615, Nov. 2022.
9. J. Shi *et al.*, "XMorpher: Full Transformer for Deformable Medical Image Registration via Cross Attention," Jun. 16, 2022, *arXiv*: arXiv:2206.07349.
10. Z. Xu *et al.*, "Evaluation of Six Registration Methods for the Human Abdomen on Clinically Acquired CT," *IEEE Trans Biomed Eng*, vol. 63, no. 8, pp. 1563–1572, Aug. 2016.
11. Z. Qin, H. Yi, Q. Lao, and K. Li, "Medical Image Understanding with Pretrained Vision Language Models: A Comprehensive Study," Feb. 08, 2023, *arXiv*: arXiv:2209.15517.
12. Z. Wang *et al.*, "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text," *Conference on Empirical Methods in Natural Language Processing*, 2022, 3876–3887.

13. S.-C. Huang *et al.*, "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 3922–3931.
14. J. Pan *et al.*, "MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning," MICCAI 2025. Lecture Notes in Computer Science, vol 15966.
15. S. Wang *et al.*, "Knowledge Graph Applications in Medical Imaging Analysis: A Scoping Review," *Health Data Sci*, vol. 2022, p. 9841548, 2022.
16. J. Zhou *et al.*, "Graph Neural Networks: A Review of Methods and Applications," *AI Open*, Volume 1, 2020, Pages 57-81, ISSN 2666-6510.
17. Y. Mou, "Knowledge Graph-Enhanced Vision-to-Language Multimodal Models for Radiology Report Generation," in *The Semantic Web: ESWC 2024 Satellite Events*, Cham: Springer Nature Switzerland, 2025, pp. 115–124.
18. B. Qi *et al.*, "GREN: Graph-Regularized Embedding Network for Weakly-Supervised Disease Localization in X-Ray Images," *IEEE J Biomed Health Inform*, vol. 26, no. 10, pp. 5142–5153, Oct. 2022.
19. X. Yu *et al.*, "CGNet: A graph-knowledge embedded convolutional neural network for detection of pneumonia," *Inf Process Manag*, vol. 58, no. 1, p. 102411, Jan. 2021.
20. X. Fu *et al.*, "Graph-Based Intercategory and Intermodality Network for Multilabel Classification and Melanoma Diagnosis of Skin Lesions in Dermoscopy and Clinical Images," *IEEE Trans Med Imaging*, vol. 41, no. 11, pp. 3266–3277, Nov. 2022.
21. D. S. Marcus *et al.*, "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *J Cogn Neurosci*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007.
22. "IXI Dataset – Brain Development." Available: https://brain-development.org/ixi-dataset/
23. D. W. Shattuck *et al.*, "Construction of a 3D probabilistic atlas of human cortical structures," *Neuroimage*, vol. 39, no. 3, pp. 1064–1080, Feb. 2008.
24. X. Jia *et al.*, "U-Net vs Transformer: Is U-Net Outdated in Medical Image Registration?" in *Machine Learning in Medical Imaging*, C. Lian, X. Cao, I. Rekik, X. Xu, and Z. Cui, Eds., Cham: Springer Nature Switzerland, 2022, pp. 151–160.
25. E. Perez *et al.*, "FiLM: Visual Reasoning with a General Conditioning Layer," *AAAI Press*, Article 483, 3942–3951.
26. A. Radford *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," I*nternational conference on machine learning*, pp. 8748–8763, PMLR, 2021.
27. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," Feb. 06, 2018, arXiv: arXiv:1710.10903.
28. B. Fischl, "Freesurfer," Neuroimage, vol. 62, no. 2, pp. 774–781, 2012
29. A. V. Dalca, J. Guttag, and M. R. Sabuncu, "Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 9290–9299.
30. A. Hering *et al.*, "Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning," *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 697–712, Mar. 2023.
31. X. Jia *et al.*, "Fourier-Net: Fast Image Registration with Band-limited Deformation," In *AAAI Press*, Vol. 37, Article 113, 1015–1023.
32. Y. Liu, L. Zuo, S. Han, Y. Xue, J. L. Prince, and A. Carass, "Coordinate translator for learning deformable medical image registration," in *International Workshop on Multiscale Multimodal Medical Imaging*, pp. 98–109, Springer, 2022.

33. X. Cheng, T. Zhang, W. Lu, Q. Meng, A. F. Frangi, and J. Duan, "SACB-Net: Spatial-awareness Convolutions for Medical Image Registration," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2025, pp. 5227-5237.
34. M. Modat *et al.*, "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine*, vol. 98, no. 3, pp. 278–284, Jun. 2010.
35. J. Chen *et al.*, "Deformer: Towards Displacement Field Learning for Unsupervised Medical Image Registration," in *MICCAI 2022*. Lecture Notes in Computer Science, vol 13436.
36. X. Chen *et al.*, "Spatially Covariant Image Registration with Text Prompts," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 7, pp. 12925-12936, July 2025.