Diffusion-Guided Anatomical Position Encoding for Dense Longitudinal CT Correspondence

Mingrui Zhuang[1,2], SongYao Zhang[1,2], Qinhe Zhang[3], Rui Xu[4], Ailian Liu[3], Xin Fan[4], and Hongkai Wang[1,2,5(✉)]

[1] Central Hospital of Dalian University of Technology (Dalian Municipal Central Hospital), Dalian 116033, China
[2] School of Biomedical Engineering, Faculty of Medicine, Dalian University of Technology, Dalian 116024, China
wang.hongkai@dlut.edu.cn
[3] The First Affiliated Hospital of Dalian Medical University, Dalian 116000, China
[4] School of Software, Dalian University of Technology, Dalian 116024, China
[5] Dalian Key Laboratory of Intelligent Health and Digital Twin, Dalian, China

Abstract. Dense anatomical correspondence across computed tomography (CT) scans is essential for longitudinal disease analysis, inter-phase alignment, and quantitative assessment. However, establishing reliable correspondence in longitudinal CT remains challenging: scan pairs are relatively scarce and temporally consistent annotations are costly to obtain, limiting the supervision for learning-based methods. Moreover, substantial appearance variations caused by respiration, pathology, and acquisition differences further challenge explicit registration and hinder generalization. In this work, we propose a diffusion-guided anatomical position encoding framework that leverages large-scale cross-subject CT data to learn generalized anatomical consistency and transfer it to intra-patient longitudinal tracking. Intermediate features from a pretrained CT generative diffusion model are leveraged as anatomical priors to supervise voxel-wise position embeddings, enabling correspondence learning without spatial alignment or task-specific annotations. To promote globally consistent and anatomically discriminative matching, we introduce dual momentum contrastive dictionaries for body-wide feature regularization, followed by a coarse-to-fine refinement network for sub-voxel correspondence precision. We evaluate the proposed method on lung landmark tracking in the DIR-Lab 4DCT dataset and multi-organ lesion tracking in the Deep Longitudinal Study (DLS) dataset. On DIR-Lab, our method achieves the lowest average TRE of 1.44 mm; on DLS, it attains a MED of 5.7 mm and a CPM@10mm of 87.63%, outperforming state-of-the-art unsupervised methods and approaching supervised trackers without using lesion annotations during training. These results highlight the potential of diffusion-guided anatomical priors for accurate and annotation-efficient longitudinal CT correspondence. The code is available at: https://github.com/DlutMedimgGroup/diffusion-ct-tracking.

Keywords: Dense anatomical correspondence · Representation learning · Longitudinal image analysis.

1 Introduction

Computed tomography (CT) is widely used for longitudinal clinical assessment, where consistent dense anatomical correspondence across scans is essential for disease monitoring (e.g., tracking tumor growth or treatment response) [12, 4]. Yet correspondence remains challenging: even within the same patient, scans acquired at different time points can vary substantially due to respiration, pathology, and acquisition factors, causing large appearance and geometric changes [26]. In addition, temporally consistent longitudinal annotations are scarce and costly, limiting supervised learning. Robust tracking therefore requires disentangling stable anatomy from scan-specific intensity and deformation variations.

Existing longitudinal CT tracking approaches can be broadly categorized into supervised tracking, registration-based alignment, and temporal or motion-driven strategies. Supervised methods rely on annotated landmarks or voxel labels and perform well with sufficient supervision [20, 5], but their dependence on task-specific annotations limits scalability. Registration-based methods align images via estimated transformations [7, 17, 18, 13], yet often degrade under large anatomical variability or disease-induced changes. Temporal strategies exploit short-term continuity [15], which is less applicable to sparsely acquired longitudinal CT. These limitations motivate anatomically grounded formulations that avoid heavy reliance on dense supervision or explicit alignment.

Recent studies have explored implicit dense anatomical correspondence via representation learning [25, 23, 2, 21], where voxels are embedded into a feature space and matched by similarity. Supervision typically relies on within-scan or within-subject cues, such as overlapping patches [25], organ-mask-derived anatomical supervision [2], data augmentations [23], or spatial proximity constraints [10, 21]. While effective for modeling intra-subject consistency, these signals provide limited guidance for learning anatomically consistent representations that generalize across subjects and time. Moreover, Euclidean distance is often used as a proxy for anatomical similarity, whose reliability varies with body size, organ morphology, and pathology. These limitations motivate complementary supervision that captures stable cross-subject anatomical regularities for longitudinal tracking.

Motivated by this gap, we turn to generative diffusion models, which have demonstrated strong capability in modeling anatomical variability in high-fidelity CT synthesis [11, 22]. To generate anatomically plausible images across subjects, such models must recover both anatomical identity and spatial context during denoising, resulting in intermediate representations that implicitly encode consistent anatomical organization [19]. Building on this observation, we reinterpret pretrained diffusion models as carriers of cross-subject anatomical knowledge and propose a diffusion-guided anatomical position encoding framework that distills the similarity structure of this anatomical prior into dense voxel-wise correspondence representations. This guidance enables correspondence that generalizes across scans while remaining robust to appearance variability, without requiring explicit alignment or task-specific annotations.

Our contributions are two-fold. **First**, we introduce a diffusion-guided anatomical position encoding paradigm that transfers cross-subject structural knowledge from a pretrained generative model to dense CT correspondence learning without explicit alignment or task-specific annotations. **Second**, we improve correspondence robustness through globally consistent feature discrimination and a coarse-to-fine refinement strategy for precise local matching. We validate the method on two longitudinal scenarios: lung landmark tracking in 4DCT and multi-organ lesion tracking in DeepLesion, demonstrating its effectiveness for real-world disease monitoring.

2 Methods

We propose a diffusion-guided two-stage framework for dense CT tracking. In Stage 1, voxel-level position embeddings are learned by distilling intermediate features from a pretrained CT diffusion model (Fig. 1). In Stage 2, coarse correspondences are refined using local image context (Fig. 2). At inference, Stage 1 provides global anatomical localization, while Stage 2 performs local refinement for sub-voxel precision.

2.1 Diffusion-Guided Anatomical Position Encoding

We learn anatomically informed dense position embeddings by distilling structural knowledge from a pretrained CT latent diffusion model. We adopt the MAISI backbone [11], trained on large-scale CT volumes. Features are extracted from the ResNet layers of the first and third decoder blocks of the diffusion U-Net [11]. We empirically found that this combination provides the best correspondence accuracy, as shallow decoder features preserve fine-grained anatomical detail while deeper features encode broader anatomical context. The framework is architecture-agnostic and can extend to other pretrained generative backbones.

During training, two CT volumes I_A and I_B are randomly sampled, and partially overlapping patch pairs are extracted. From each overlapping region, n spatial locations are sampled as supervision queries.

The diffusion backbone is frozen, while a lightweight 3D decoder (adapter) is trained to map intermediate denoising features to spatially aligned guidance features. Given a patch P, we feed it into the frozen latent diffusion model and extract intermediate features from the denoising U-Net at a fixed timestep of the reverse process. These internal activations are then passed through the adapter to obtain spatially aligned guidance features $\tilde{\mathbf{f}}$. In parallel, our primary trainable component, an anatomical position embedding network, produces voxel-level embeddings $\mathbf{z}(x) \in \mathbb{R}^d$. This network adopts a 3D U-Net architecture with a transformer bottleneck to capture both local spatial detail and global anatomical context.

Rather than matching individual feature vectors, we align the similarity structure induced by the adapted diffusion features and the learned embeddings. For each of the n sampled queries, similarities are computed against a set

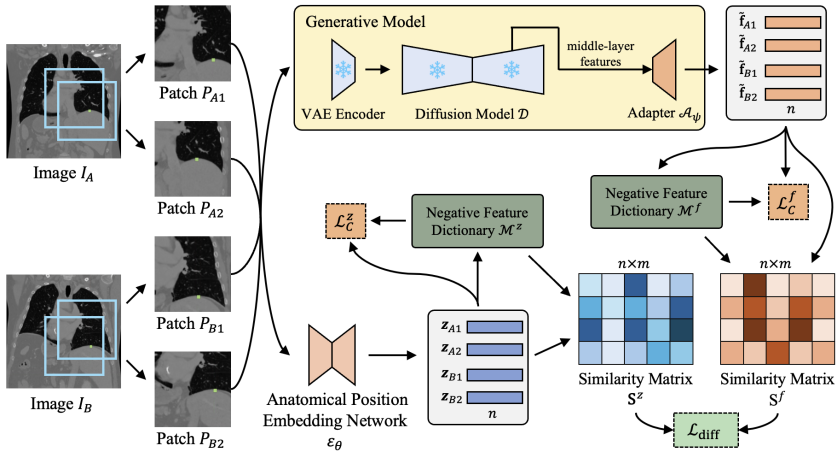

Fig. 1. Stage 1: Training framework for diffusion-guided anatomical position encoding. A frozen diffusion backbone with a trainable adapter provides anatomical guidance, while a learnable encoder generates voxel-level embeddings. Dual momentum dictionaries enforce global anatomical discrimination, and structural similarity is aligned via KL divergence.

of keys (Sec.2.2), forming similarity distributions. After row-wise normalization, we minimize the Kullback–Leibler divergence

$$\mathcal{L}_{\text{diff}} = \text{KL}(\mathbf{S}^f \,\|\, \mathbf{S}^z),$$

encouraging the embedding network to match the similarity structure induced by the diffusion features. In this way, the diffusion model acts as a structural teacher and the embedding network is optimized end-to-end via $\mathcal{L}_{\text{diff}}$ to capture anatomically consistent spatial relationships.

2.2 Global Anatomical Discrimination with Dual Momentum Dictionaries

While diffusion-guided structural alignment transfers anatomical similarity, additional regularization is required to ensure global discrimination across distant anatomical regions. To this end, we maintain two independent momentum-updated feature dictionaries: one for the adapted diffusion features and one for the learned embeddings.

Each dictionary stores m feature vectors collected from diverse spatial locations across training volumes, providing a large body-wide negative pool for contrastive learning. For the n sampled query locations in Sec.2.1, similarities are computed against all m entries in the corresponding dictionary, forming similarity matrices $\mathbf{S}^f, \mathbf{S}^z \in \mathbb{R}^{n \times m}$.

We apply an InfoNCE-style contrastive loss separately to both branches, $\mathcal{L}_C^f$ and $\mathcal{L}_C^z$, encouraging global anatomical separability in both the adapted

diffusion feature space and the embedding space. Each dictionary is updated using a momentum encoder to stabilize feature evolution.

The overall objective is defined as

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda(\mathcal{L}_C^f + \mathcal{L}_C^z),$$

combining structural alignment and global discrimination to produce anatomically coherent and globally separable position embeddings.

2.3 Coarse-to-Fine Refinement for Dense Correspondence

After learning globally discriminative embeddings, dense correspondence is estimated in a coarse-to-fine manner (Fig. 2). In the coarse stage, embeddings from the reference volume I_r and target volume I_t are matched to obtain an initial correspondence map that preserves global anatomical consistency. Stage 2 then refines each coarse match within a local neighborhood, where the residual correspondence ambiguity is substantially reduced, enabling more precise spatial localization.

To improve spatial precision, a refinement network operates on local patches centered at coarse matches. During training, overlapping patch pairs from the two volumes are sampled and optimized with an InfoNCE objective, where spatially aligned locations within the overlapping region serve as positives and other locations as negatives. This self-supervised design enables fine-grained local matching without ground-truth deformation or landmark annotations. Each patch is processed by a lightweight local refinement encoder, followed by four transformer blocks for intra-patch and inter-image context modeling, and finally upsampled by a refinement decoder to the original resolution.

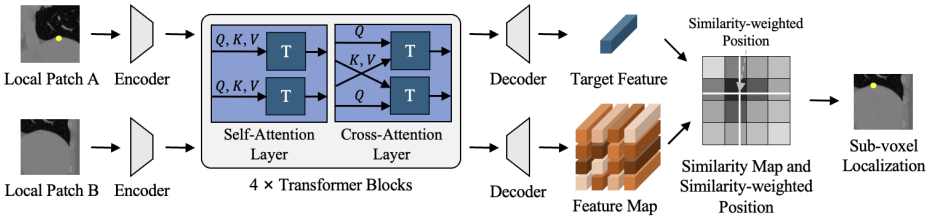

Fig. 2. Stage 2: Refinement network for sub-voxel alignment. Given locally matched reference and target patches, the network predicts refined correspondence within a search window. Final coordinates are obtained via similarity-weighted averaging, enabling sub-voxel precision.

For each query, similarity within the local search window is computed, and the final coordinate is obtained via similarity-weighted averaging of candidate voxel coordinates. By combining global embedding-based matching with local

transformer refinement, the framework achieves anatomically consistent, sub-voxel accurate tracking without explicit deformation estimation or prior registration.

3 Experiments and Results

3.1 Datasets and Implementation Details

The anatomical position encoder is trained in an unsupervised manner using FLARE2023 [14] (4,000 abdominal CT scans) and LIDC-IDRI [1] (833 thoracic CT scans), without landmark or correspondence annotations.

Evaluation is conducted on two benchmarks. The DIR-Lab 4DCT dataset contains 10 thoracic cases with annotated inspiration–expiration lung landmarks. The Deep Longitudinal Study (DLS) dataset [5], a subset of DeepLesion, includes 3,891 longitudinal lesion pairs across multiple organs with point-level annotations; results are reported on the official 480 testing pairs.

Tracking performance is measured using TRE, MED, and CPM. All volumes are resampled to $1\,\mathrm{mm}^3$ isotropic resolution, with intensities clipped to [-1000, 1000] HU and normalized to [0,1]. During training, overlapping patch pairs are sampled from two volumes, with $n = 512$ queries per region compared against a momentum dictionary of size $m = 65{,}536$. Global patches of size 192^3 are embedded into $d = 32$-dimensional vectors, and Stage 2 additionally samples local refinement patches of size 64^3 around coarse matches. The diffusion backbone is frozen. Models are trained for 50 epochs using AdamW (lr 3×10^{-4}, batch size 2), with temperature $\tau = 0.2$, loss weight $\lambda = 0.01$, and momentum 0.999.

3.2 Qualitative Analysis of Diffusion-Guided Correspondence

Fig. 3 visualizes correspondence localization using raw diffusion features and the learned anatomical position embeddings. For each example, a query location is defined in the reference image, and correspondence scores are computed on the target image via similarity-based matching. Raw diffusion feature matching is performed by directly computing cosine similarity between intermediate diffusion features at the query location and all spatial locations in the target image.

Diffusion matching yields anatomically coherent but coarse responses, while the distilled embeddings produce more concentrated maps, indicating sharpened correspondence after distillation and contrastive learning.

3.3 Correspondence Accuracy

4DCT Results. We evaluate our method on the DIR-Lab 4DCT dataset, which contains ten inspiration–expiration CT pairs with annotated landmark correspondences under substantial respiratory deformation.

Table 1 reports case-wise TRE comparisons with classical and learning-based registration methods. Classical optimization-based methods (Elastix [16],

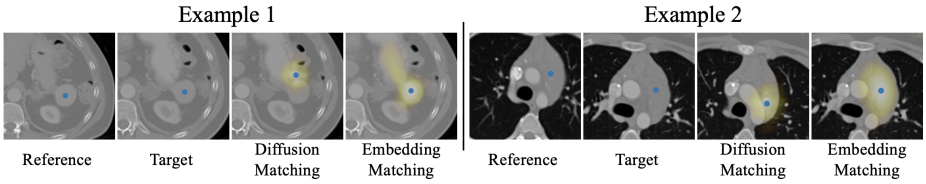

Fig. 3. Qualitative comparison of correspondence localization. A query location (blue dot) is selected in the reference image. Heatmaps on the target image are obtained by (i) diffusion-feature matching and (ii) the proposed position-embedding matching. Examples are drawn from the DLS test set.

Table 1. Case-wise target registration error (TRE, mm) on the DIR-Lab 4DCT dataset. C1–C10 denote the ten individual 4DCT cases provided in the benchmark. Smaller values indicate better registration accuracy.

Method	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Avg.
Initial	3.89	4.34	6.94	9.83	7.48	10.89	11.03	14.99	7.92	7.30	8.46
Elastix [16]	1.19	1.12	1.53	1.97	2.40	2.68	2.09	2.50	1.67	2.32	1.95
ANTs	1.17	1.09	1.38	1.67	2.21	2.39	1.97	2.61	2.10	1.82	1.84
TransMorph [6]	1.18	1.16	1.67	2.04	1.98	2.74	2.31	4.64	2.03	2.28	2.20
SDHNet [28]	1.14	1.13	1.36	1.71	1.77	1.99	1.86	2.92	1.72	1.87	1.75
LungRegNet [9]	1.18	1.13	1.41	1.93	1.83	2.11	1.76	2.62	1.69	**1.53**	1.72
MANet [27]	**1.09**	**1.01**	**1.24**	1.63	1.66	2.10	1.61	2.33	1.47	1.54	1.57
ADC-Net [24]	1.18	1.22	1.40	1.60	1.67	1.75	1.58	1.95	**1.41**	1.54	1.53
Ours	1.17	1.21	1.29	**1.41**	**1.43**	**1.47**	**1.57**	**1.57**	1.62	1.66	**1.44**

ANTs) estimate dense deformation fields via iterative registration, while recent learning-based approaches (e.g., MANet [27], ADC-Net [24]) learn correspondence through data-driven motion modeling. Our approach achieves the lowest average TRE of 1.44 mm among all compared methods, outperforming recent unsupervised methods such as ADC-Net (1.53 mm) and MANet, and reducing TRE by over 0.4 mm compared with classical baselines.

DeepLesion Results. We further evaluate our method on the Deep Longitudinal Study (DLS) dataset [5], which contains 3,891 longitudinal lesion pairs with annotated lesion centers across multiple organs. This task involves identifying corresponding lesions under acquisition and disease variations.

Table 2 compares our method with supervised tracking models and unsupervised registration-based approaches. Among unsupervised methods, our approach achieves the best performance, reducing MED to 5.7 mm (vs. 5.9 mm for MS-SSL-AE [21] and 6.5 mm for SAM [25]) and attaining the highest CPM@10mm (87.63%). Notably, it approaches the accuracy of the supervised TLT model [20] (MED 6.0 mm) without using lesion annotations during training.

Compared with classical registration methods (Affine [16] and VoxelMorph [3]), the improvement demonstrates that anatomically informed embeddings provide stronger cross-scan consistency than purely spatial alignment.

Table 2. Comparison of longitudinal lesion tracking performance on the Deep Longitudinal Study (DLS) dataset. CPM denotes correct prediction metric, and MED denotes mean Euclidean distance (mm).

Method	CPM@ 10 mm	CPM@ Radius	MED_x (mm)	MED_y (mm)	MED_z (mm)	MED (mm)
Supervised						
DLT-Mix [5]	78.65	88.75	3.1	3.1	4.2	7.1
DLT [5]	78.85	86.88	3.5	2.9	4.0	7.0
TransT [8]	79.59	88.99	3.4	5.4	1.8	7.6
TLT [20]	87.37	**95.32**	3.0	3.7	**1.7**	6.0
Unsupervised						
Affine [16]	48.33	65.21	4.1	5.4	7.1	11.2
VoxelMorph [3]	49.90	65.59	4.6	5.2	6.6	10.9
SAM [25]	81.67	90.21	2.9	2.5	3.6	6.5
MS-SSL-AE [21]	83.13	91.87	2.9	**2.2**	3.1	5.9
Ours	**87.63**	93.28	**2.4**	2.4	3.6	**5.7**

Table 3. Ablation study on the Deep Longitudinal Study (DLS) dataset. We evaluate the contribution of diffusion-guided supervision, dual momentum feature dictionaries, and the Stage 2 refinement network for longitudinal lesion tracking. Lower MED and higher CPM indicate better performance.

Diffusion-Guided Supervision	Dual Momentum Dictionaries	Stage 2 Refinement	MED (mm)	CPM@10 mm
✓	×	×	7.22	80.38
×	✓	×	7.45	81.99
✓	✓	×	6.69	84.41
×	✓	✓	6.16	86.29
✓	✓	✓	**5.67**	**87.63**

3.4 Ablation Study

We conduct ablation experiments on the DLS dataset to evaluate the contributions of diffusion-guided supervision, dual momentum dictionaries, and the Stage 2 refinement network (Table 3).

Diffusion-guided supervision alone provides effective anatomical guidance, while dual momentum dictionaries improve global discrimination. Combining both further enhances tracking accuracy, demonstrating their complementary roles. The Stage 2 refinement network yields additional gains by correcting local correspondence errors.

4 Conclusion

We present a diffusion-feature-guided anatomical position encoding framework for dense longitudinal CT correspondence without explicit alignment or corre-

spondence annotations. The learned embeddings capture global anatomical consistency, while a coarse-to-fine refinement module further improves local matching to sub-voxel precision. Experiments on 4DCT and DeepLesion demonstrate strong performance for robust and annotation-efficient longitudinal correspondence analysis. Future work will investigate uncertainty-aware correspondence estimation and no-match detection for newly appearing or disappearing lesions and severe anatomical changes.

Acknowledgments. This study was supported in part by the China Postdoctoral Science Foundation (No. 2025M772947), the Young Scientists Fund of the National Natural Science Foundation of China (No. 62403103), the Joint Funds of the National Natural Science Foundation of China (No. U24A20753), the general program of National Natural Science Fund of China (No. 81971693), the Fundamental Research Funds for the Central Universities (DUT24RC(3)050), the funding of Liaoning Key Lab of IC & BME System and Dalian Engineering Research Center for Artificial Intelligence in Medical Imaging.

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

References

1. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
2. Bai, X., Bai, F., Huo, X., Ge, J., Lu, J., Ye, X., Shu, M., Yan, K., Xia, Y.: Uae: Universal anatomical embedding on multi-modality medical images. Medical Image Analysis **103**, 103562 (2025)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
4. Bartolomucci, A., Nobrega, M., Ferrier, T., Dickinson, K., Kaorey, N., Nadeau, A., Castillo, A., Burnier, J.V.: Circulating tumor dna to monitor treatment response in solid tumors and advance precision oncology. NPJ Precision Oncology **9**(1), 84 (2025)
5. Cai, J., Tang, Y., Yan, K., Harrison, A.P., Xiao, J., Lin, G., Lu, L.: Deep lesion tracker: monitoring lesions in 4d longitudinal imaging studies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15159–15169 (2021)
6. 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)
7. Chen, J., Liu, Y., Wei, S., Bian, Z., Subramanian, S., Carass, A., Prince, J.L., Du, Y.: A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond. Medical Image Analysis **100**, 103385 (2025)

8. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8126–8135 (2021)
9. Fu, Y., Lei, Y., Wang, T., Higgins, K., Bradley, J.D., Curran, W.J., Liu, T., Yang, X.: Lungregnet: an unsupervised deformable image registration method for 4d-ct lung. Medical physics **47**(4), 1763–1774 (2020)
10. Goncharov, M., Samokhin, V., Soboleva, E., Sokolov, R., Shirokikh, B., Belyaev, M., Kurmukov, A., Oseledets, I.: Anatomical positional embeddings. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 68–77. Springer (2024)
11. Guo, P., Zhao, C., Yang, D., Xu, Z., Nath, V., Tang, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., et al.: Maisi: Medical ai for synthetic imaging. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4430–4441. IEEE (2025)
12. Jin, C., Yu, H., Ke, J., Ding, P., Yi, Y., Jiang, X., Duan, X., Tang, J., Chang, D.T., Wu, X., et al.: Predicting treatment response from longitudinal images using multi-task deep learning. Nature communications **12**(1), 1851 (2021)
13. Khor, H.G., Ning, G., Sun, Y., Lu, X., Zhang, X., Liao, H.: Anatomically constrained and attention-guided deep feature fusion for joint segmentation and deformable medical image registration. Medical Image Analysis **88**, 102811 (2023)
14. Ma, J., Zhang, Y., Gu, S., Ge, C., Mae, S., Young, A., Zhu, C., Yang, X., Meng, K., Huang, Z., et al.: Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the flare22 challenge. The Lancet Digital Health **6**(11), e815–e826 (2024)
15. Ma, M., Zhang, X., Li, Y., Wang, X., Zhang, R., Wang, Y., Sun, P., Wang, X., Sun, X.: Convlstm coordinated longitudinal transformer under spatio-temporal features for tumor growth prediction. Computers in Biology and Medicine **164**, 107313 (2023)
16. Marstal, K., Berendsen, F., Staring, M., Klein, S.: Simpleelastix: A user-friendly, multi-lingual library for medical image registration. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 134–142 (2016)
17. 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)
18. Su, H., Yang, X.: Nonuniformly spaced control points based on variational cardiac image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 634–644. Springer (2023)
19. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in Neural Information Processing Systems **36**, 1363–1389 (2023)
20. Tang, W., Kang, H., Zhang, H., Yu, P., Arnold, C.W., Zhang, R.: Transformer lesion tracker. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 196–206. Springer (2022)
21. Vizitiu, A., Mohaiu, A.T., Popdan, I.M., Balachandran, A., Ghesu, F.C., Comaniciu, D.: Multi-scale self-supervised learning for longitudinal lesion tracking with optional supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 573–582. Springer (2023)
22. Wang, H., Liu, Z., Sun, K., Wang, X., Shen, D., Cui, Z.: 3d meddiffusion: A 3d medical latent diffusion model for controllable and high-quality medical image generation. IEEE Transactions on Medical Imaging (2025)

23. 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)
24. Wu, W., Gao, Y., Jin, X., Zhang, R., Pan, Y., Gao, X.: An unsupervised anatomy-aware dual-constraint cascade network for lung computed tomography deformable image registration. Engineering Applications of Artificial Intelligence **158**, 111548 (2025)
25. 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)
26. Yang, C., Huang, L., Sucharit, W., Xie, H., Huang, X., Li, Y.: Transformer-enhanced vertebrae segmentation and anatomical variation recognition from ct images. Scientific Reports **15**(1), 34329 (2025)
27. Yang, J., Yang, J., Zhao, F., Zhang, W.: An unsupervised multi-scale framework with attention-based network (manet) for lung 4d-ct registration. Physics in Medicine & Biology **66**(13), 135008 (2021)
28. Zhou, S., Hu, B., Xiong, Z., Wu, F.: Self-distilled hierarchical network for unsupervised deformable image registration. IEEE Transactions on Medical Imaging **42**(8), 2162–2175 (2023)