Guiding Registration with Emergent Similarity from Pre-Trained Diffusion Models

Nurislam Tursynbek[1], Hastings Greer[1], Başar Demir[1], Marc Niethammer[2]

[1]UNC Chapel Hill, [2]UCSD,   Correspondence: `nurislam@cs.unc.edu`

Abstract. Diffusion models, while trained for image generation, have emerged as powerful foundational feature extractors for downstream tasks. We find that off-the-shelf diffusion models, trained exclusively to generate natural RGB images, can identify semantically meaningful correspondences in medical images. Building on this observation, we propose to leverage diffusion model features as a similarity measure to guide deformable image registration networks. We show that common intensity-based similarity losses often fail in challenging scenarios, such as when certain anatomies are visible in one image but absent in another, leading to anatomically inaccurate alignments. In contrast, our method identifies true semantic correspondences, aligning meaningful structures while disregarding those not present across images. We demonstrate superior performance of our approach on two tasks: multimodal 2D registration (DXA to X-Ray) and monomodal 3D registration (brain-extracted to non-brain-extracted MRI). Code: https://github.com/uncbiag/dgir

1 Introduction

Deep learning based deformable image registration [44,2,11] has shown to be useful in practice by delivering fast and accurate estimates of spatial correspondences (deformation fields) to align medical images. Deep networks are trained on populations of image pairs; typically by minimizing the sum of two losses: a similarity loss between the warped moving and the original fixed images and a regularization loss to encourage smoothness of the predicted deformation field.

Conventional similarity losses (Mean Squared Error(MSE) [2], Localized Normalized Cross-Correlation (LNCC) [2], Normalized Gradient Fields (NGF) [15], Modality Independent Neighborhood Descriptors (MIND) [16]) assume identical anatomical structures in images and fail when anatomies are missing in one of the images [7]. The "missing anatomy" scenario often appears in practice both in monomodal and multimodal registration. For example, to align a DXA (Dual Energy X-Ray) scan, which focuses on bone density and ignores soft tissues, to a standard X-Ray, which visualizes soft tissue in addition to bone, the model may stretch the bones in the DXA scan to fill the soft tissue space (see Fig. 4).

In monomodal registration, challenges arise when aligning pre- and post-operative images, or images at differing processing stages, such as magnetic resonance images (MRIs) of the head with and without brain extraction. Here, registration networks often attempt to compensate for the missing structures
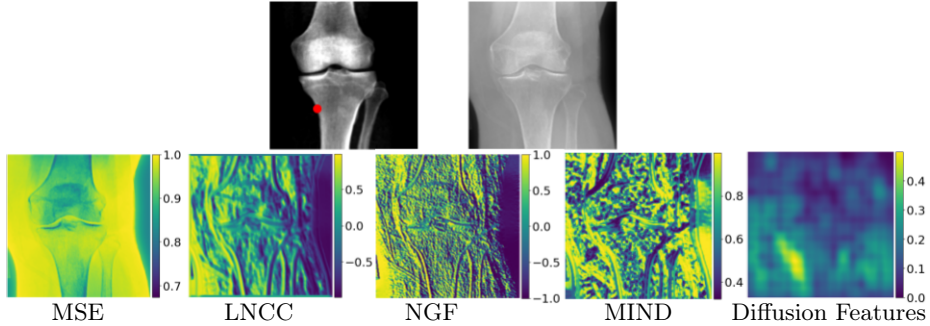
Fig. 1: **Heatmaps.** First row: DXA scan (a red dot on the boundary between the bone and the background) and X-Ray. Second row: Heatmaps for different similarity measures indicating how close each point in the X-Ray is to the corresponding red point in the DXA scan. Correspondences for conventional pixel-based similarity measures are ambiguous as many pixels show similar values, while diffusion features identify semantically meaningful correspondences. DXA images reproduced by kind permission of the UK Biobank®

by expanding existing ones, as conventional similarity losses align contrasting edges, such as those between background and outermost anatomies, instead of desired internal alignment (see Fig. 1, 4, 6). Therefore, for images with missing structures anatomy-aware losses with deep semantic understanding is required.

Diffusion models [17,31] have revolutionized computer vision with high-fidelity image generation. Beyond achieving state-of-the-art image synthesis [9,31], the impressive generative abilities of these models suggest that they capture both low-level and high-level features, highlighting their potential as general representation learners [12]. These representations have been found useful in many downstream tasks such as for classification [26,43,5], segmentation [3,42,39], editing [41,29,33], depth and geometry estimation [45,19,22].

Moreover, it has been shown that features from diffusion models, trained solely for image generation, have an emergent knowledge of semantic landmarks and correspondences even under challenging conditions such as occlusions and appearance variations [38,23,25]. Intriguingly, this semantic correspondence ability extends to medical images (see Fig. 2a), despite the significant domain gap and the absence of medical images in the training. Therefore, diffusion features may provide a useful signal for registration, especially in challenging scenarios (such as "missing anatomy"), where deep semantic understanding is necessary.

Our contributions:

- We observe that diffusion models, trained to generate natural RGB images, also capture correspondences in medical images even in multimodal scenarios or where not all anatomies are present in an image pair.
- We propose a registration similarity loss based on comparing diffusion features between a warped moving and a fixed image.
- We show superior performance on 2D registration of DXA to X-Ray and 3D registration of brain-extracted MRI to non-brain-extracted MRI.

(a) Diffusion Features.   (b) Diffusion Features + LNCC

Fig. 2: **Correspondences.** Intermediate diffusion features, resized to image resolution, are vector-valued descriptors for each pixel. For each keypoint pixel in the left (DXA) the most similar (cosine similarity) pixel in the right (X-Ray) is connected with a line. (a) Diffusion features find complex correspondences in medical images with missing anatomies, although trained on RGB images. (b) LNCC on diffusion features further improves correspondences, considering surrounding pixels. DXA images reproduced by kind permission of the UK Biobank®

2   Related work

Diffusion Models and Image Registration have been studied together in the following ways. DiffuseMorph [20] jointly trains diffusion and registration networks sequentially, passing noise predicted from the diffusion network as input to the registration network. FSDiffReg [30] extends this approach by adding intermediate diffusion features as input to the registration network. Diff-Def [36] trains a diffusion model to generate deformation fields conditioned on specific parameters, guided by a registration network to optimize atlas deformation. DiffuseReg [46] combines diffusion and registration by iteratively denoising a deformation field from noise. While the concept of jointly training diffusion and registration networks is well explored, our work is the first work to explore off-the-shelf pre-trained diffusion models to define a similarity loss for registration.

Similarity losses for image registration are well-studied for optimization-based [1] and deep-learning approaches [18,2]. Common choices for monomodal registration are MSE and LNCC [2], while for multimodal registration - MIND [16] and NGF [15]. These losses work well when images share anatomical structures, however they often fail when some anatomies are visible in one image and absent in another [7]. Seg-Guided-MMReg [7] guides multimodal registration with segmentaions as they are same across modalities. We find that images with missing anatomies not only share segmentation masks, but also have similar diffusion features. DeepSim [6] uses cosine similarity between features of a pre-trained autoencoder (or segmentation model) as a similarity loss. Unfortunately, it only works for monomodal scenarios with shared anatomical structures [6] and for each dataset, its own feature extractor needs to be trained, while diffusion models are training-free general feature extractors. Several works [35,21] used non-learning optimization-based registration on foundational self-supervised features[4,28]. We show how to incorporate knowledge from foundational diffusion models into the training of deep neural networks for registration.

3   Method

Diffusion models [17] are composed of a forward and a reverse process. The forward process is a T-step noising of a clean image $\mathbf{x}_0 \in \mathcal{R}^{H \times W \times C}$ with added Gaussian $\boldsymbol{\epsilon} \in \mathcal{R}^{H \times W \times C}$ at each step following a noise schedule $\{\beta_t \in \mathcal{R}^1\}_{t=1}^T$, slowly corrupting the image into noise $\mathbf{x}_T \in \mathcal{R}^{H \times W \times C}$. The noisy image at timestep t is $\mathbf{x}_t = \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \overline{\alpha}_t}\boldsymbol{\epsilon}$, where $\overline{\alpha}_t = \prod_{i=1}^t (1 - \beta_i) \in \mathcal{R}^1$. The T-step reverse process predicts the noise $\boldsymbol{\epsilon}$ at each timestep t with a network $\mathbf{h}$:

$$\widehat{\boldsymbol{\epsilon}} = \mathbf{h}(\mathbf{x}_t, t) = \mathbf{h}(\sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \overline{\alpha}_t}\boldsymbol{\epsilon}, t), \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{1}$$

The diffusion network $\mathbf{h}$ is trained to minimize the mean squared error (MSE) between the true and predicted noises $\|\boldsymbol{\epsilon} - \widehat{\boldsymbol{\epsilon}}\|^2$. Importantly, we do not train the diffusion model ourselves and use an off-the-shelf pre-trained diffusion model.

3.1   Diffusion-Guided Image Registration

Conventional deep learning methods train a registration network $\mathbf{f}_\theta$, with weights θ, that takes moving image $\mathbf{A}$ and the fixed image $\mathbf{B}$ as inputs and outputs deformation map $\boldsymbol{\Phi}$, which warps $\mathbf{A}$ to the space of the fixed image, $\mathbf{A} \circ \boldsymbol{\Phi}$. Training involves minimizing a loss combining *similarity* and *regularization* terms:

$$\min_\theta \ \mathcal{L}_{sim}(\mathbf{A} \circ \boldsymbol{\Phi}, \mathbf{B}) + \lambda \mathcal{L}_{reg}(\boldsymbol{\Phi}), \lambda > 0. \tag{2}$$

As noted earlier, conventional similarity losses focus on aligning strong edges, rather than desired internal structures. While this works when both images show the same anatomies, it often causes misalignment when some anatomies are missing in one image but present in the other. To address this, we propose Diffusion-Guided Image Registration (DGIR), leveraging features from pre-trained diffusion models, found to capture semantic correspondences, for image registration.

Fig. 3 shows an overview of our framework. The input (moving image $\mathbf{A}$ and fixed image $\mathbf{B}$) and output (deformation map $\boldsymbol{\Phi}$) of the registration network $\mathbf{f}_\theta$ are the same as for typical registration networks. DGIR additionally passes warped and fixed images through a diffusion-based feature extractor $\mathbf{g}$:

$$\mathbf{g}(\mathbf{x}) = \overline{\mathbf{h}}_n(\sqrt{\overline{\alpha}_t}\mathbf{x} + \sqrt{1 - \overline{\alpha}_t}\boldsymbol{\epsilon}, t), \tag{3}$$

where $\overline{\mathbf{h}}_n$ is n-th block output of the pre-trained diffusion model $\mathbf{h}$ from Eq. (1).

Instead of defining a similarity loss directly between the warped and the fixed images, we compute it on the diffusion features to promote deep semantic alignment. Note that these diffusion features are computed on the *warped* moving image and the fixed image separately. I.e., for a perfect image alignment these features would by construction be identical. We use the $1 - LNCC$ loss on the diffusion features as our similarity loss and spatial gradient penalization [2] of the displacement field $\|\nabla(\boldsymbol{\Phi} - \mathbf{id})\|_F^2$ (where $\mathbf{id}$ is the identity map) as regularization:

$$\mathcal{L} = \mathcal{L}_{sim}(\mathbf{g}(\mathbf{A} \circ \boldsymbol{\Phi}), \mathbf{g}(\mathbf{B})) + \lambda \mathcal{L}_{reg}(\boldsymbol{\Phi}) = 1 - LNCC(\mathbf{g}(\mathbf{A} \circ \boldsymbol{\Phi}), \mathbf{g}(\mathbf{B})) + \lambda \|\nabla(\boldsymbol{\Phi} - \mathbf{id})\|_F^2, \tag{4}$$
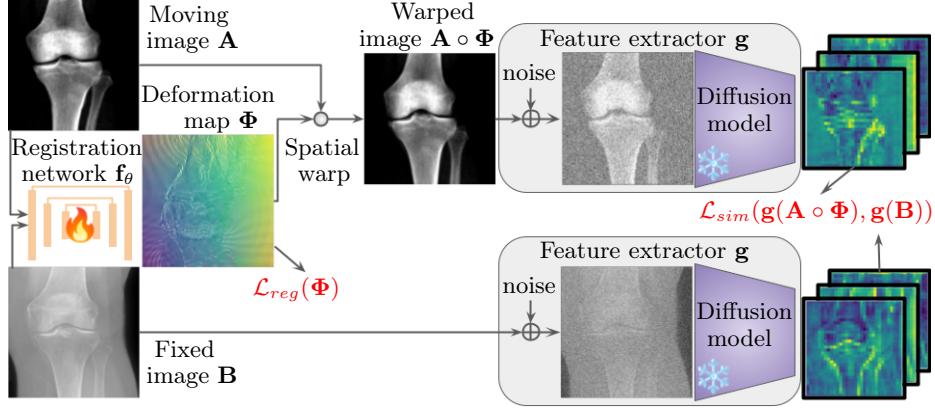
Fig. 3: **Overview of the method.** The registration network $\mathbf{f}_\theta$ is trained with a similarity loss that encourages diffusion features $\mathbf{g}(\mathbf{A} \circ \mathbf{\Phi})$ of the warped image $\mathbf{A} \circ \mathbf{\Phi}$ to be similar to the diffusion features $\mathbf{g}(\mathbf{B})$ of the fixed image $\mathbf{B}$. A regularization loss is applied to the deformation map $\mathbf{\Phi}$ for spatial smoothness. DXA images reproduced by kind permission of the UK Biobank®

3.2   Datasets and Implementation Details

Registration. For all registration networks, we use a multi-step U-Net [13] and train with Adam and learning rate $1e{-}4$. Models are implemented using PyTorch and trained on one Nvidia A6000 GPU. We set $\lambda = 1$ for all experiments.

Diffusion Models. We use a publicly available unconditional diffusion model from [9] pre-trained on $256 \times 256 \times 3$ RGB images from ImageNet [8]. This model is a U-Net [32] with 37 blocks (18 Encoder, 1 Middle, and 18 Decoder blocks).

2D experiments. We use knee DXA scans from the UK Biobank [37] and X-Rays from The Osteoarthritis Initiative (OAI) [27]. Following [7], we resize the images to 256×256 and affine aligned them to a random reference image using the affine layers of [14]. We use 98 DXA scans and 758 X-Rays as a training set and 25 DXA scans and 100 X-Rays as a test set (2500 test pairs) and manual segmentations of the 3 major knee bones (femur, tibia, and fibula) for evaluation.

3D experiments. Although the diffusion model is pre-trained on 2D RGB images, it provides useful signals for *volumetric* 3D registration. For that, we randomly select N coronal, or sagittal, or axial slices for both fixed and warped 3D images. Then, we extract diffusion features for these slices and apply the LNCC loss on them. Throughout the training, this process will cover the entire volume thereby using a 2D feature extractor for 3D registration. We use 3D brain MRIs from Neurite-OASIS [24,18], which consists of 414 3D images affinely alligned to FreeSurfer Talairach atlas [10] and center-cropped to size $160 \times 192 \times 224$. The dataset comes in two forms: raw (non-brain-extracted) MRIs and brain-extracted/skull-stripped MRIs. We use 300 images as a training set, and 114 images as a test set ($114 \times 113 = 12882$ test pairs) and 4 brain region segmentations (Cortex, White Matter, Gray Matter, CSF) for evaluation.
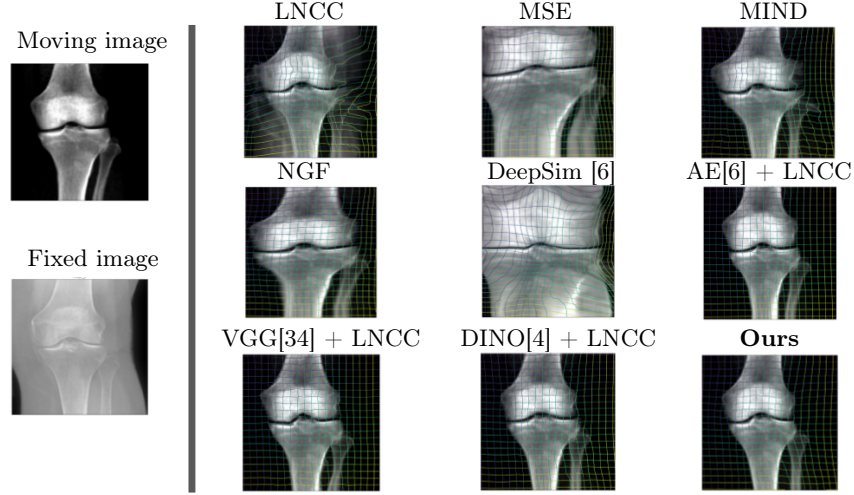
Fig. 4: **2D results.** Pixel-based similarity losses fail to capture true correspondences in the "missing anatomy" scenario. Images on the right show the warped moving image which should ideally resemble the fixed image. Our method (Diffusion Features + LNCC) performs the best, as quantitatively shown in Tab. 1. DXA images reproduced by kind permission of the UK Biobank[®]

## 4   Multimodal (DXA→X-Ray) 2D Knee Registration

We show registration results in Fig. 4 and report Dice scores of bone segmentations and percentage of pixels with negative Jacobian for the registration map, Φ, in Tab. 1. Conventional methods fail in finding appropriate semantic correspondences in the "missing anatomy" setting. They stretch the bones and fill up the space of soft tissue, which is visible in X-Rays and is missing in DXA scans.

Although other feature extractors (Autoencoder [6], VGG [34], DINO [4]) do not have this issue, the features from off-the-shelf pretrained diffusion models show the best performance (Tab. 1). The combination of LNCC with diffusion features outperforms all other baselines and even the existing state-of-the-art method Seg-Guided-MMReg [7], which was trained with segmentations.

The off-the-shelf pre-trained diffusion model from [9] uses a U-Net with 37 blocks and $T = 1000$ noise levels (timesteps). We ablated outputs of which block and what timestep to use to guide a registration network in Fig. 5. In Fig. 5a we fix the timestep at $t = 60$ and measure the average Dice scores for the networks guided with different diffusion blocks. Mid-resolution blocks (8-11 and 29-32) perform better compared to others and decoder blocks are slightly better than encoder blocks. In Fig. 5b, we fix the block number $n = 29$ and measure the average Dice scores for the networks guided with different noise timesteps. The best performance is observed when we add small-to-medium ($t \in [10; 150]$) noise. Test dice scores for our method reported in Tab. 1 are based on the model with best parameters (timestep $t = 50$, block $n = 29$) on small validation set.

Table 1: **Quantitative 2D knee registration results.** Dice scores of multi-modal registration with missing anatomy between DXA scans and X-Rays.

Method	Femur	Tibia	Fibula	Average	%|J|
LNCC [2]	0.8209	0.8300	0.7323	0.7944	0.00%
MSE [2]	0.6267	0.6938	0.4733	0.5980	0.00%
MIND [16]	0.9267	0.9407	0.8236	0.8970	0.00%
NGF [15]	0.8162	0.8350	0.4724	0.7079	0.00%
DeepSim [6]	0.6808	0.6452	0.0868	0.4709	0.00%
AE [6] + LNCC	0.9423	0.9490	0.7582	0.8832	0.00%
VGG [34] + LNCC	0.9446	0.9368	0.7207	0.8674	0.00%
DINO [4] + LNCC	0.9621	0.9584	0.8603	0.9269	0.00%
Diffusion + LNCC(Ours)	**0.9804**	**0.9739**	**0.9291**	**0.9611**	**0.00%**
Seg-Guided-MMReg [7]	0.9701	0.9684	0.8882	0.9422	0.00%
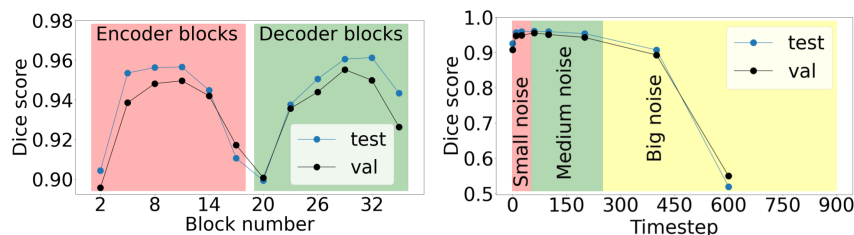
(a) Mid-resolution features are the best to guide registration. Decoder features are slightly better than encoder ones.

(b) Adding small-to-medium noise before passing an image to a pretrained diffusion model works best.

Fig. 5: **Block number and Noise timestep ablations on 2D knee dataset.**

5   3D brain registration

We leverage representations from 2D diffusion to guide 3D registration with missing anatomy (for brain MRIs: e.g., skull, neck). We fix timestep $t = 50$ and $n = 11$ block, which is comparable to the corresponding decoder block (see Fig. 5a), while significantly saves time and memory, essential for 3D registrations.

Our goal is not to push the State-Of-The-Art (SOTA) in specific registration but rather to investigate the broader utility of our loss function in a general context. We compare to two networks trained with 3D LNCC [2]: Brain-Extracted moving to Non-Brain-Extracted fixed images (BE→NBE) and Brain-Extracted moving to Brain-Extracted fixed images (BE→BE) We also compare to foundational registration method, UniGradICON [40], which was trained on diverse and extensive collection of medical images with LNCC similairty. Our network is Brain-Extracted moving to Non-Brain-Extracted fixed images (BE→NBE) trained with diffusion guidance (see Eq. (1) and Fig. 3 in Section 3.1).

Table 2: **3D results.** Our approach performs on par with the LNCC loss when moving and fixed images share anatomies and foundationl UniGradICON, while it performs significantly better when one of the images has missing anatomy.

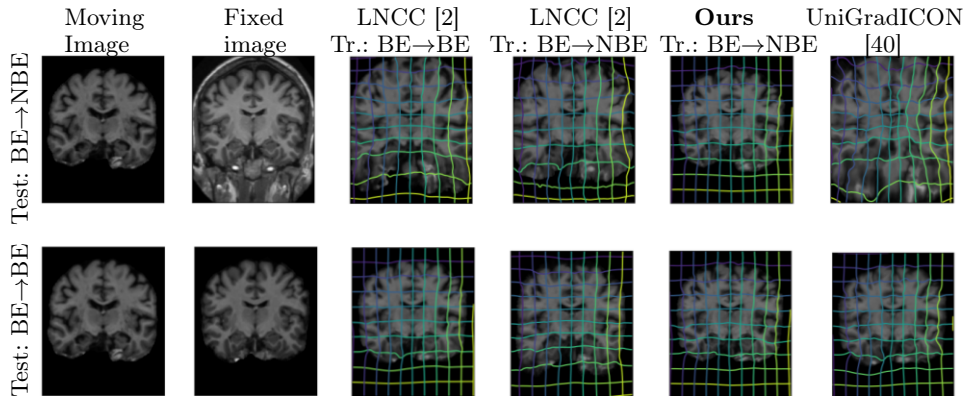| Method	Dice Score BE→NBE test	%|J| BE→NBE test	Dice Score BE→BE test	%|J| BE→BE test
LNCC [2] (Tr.:BE→BE)	0.6325	1.411%	0.8156	0.020%
LNCC [2] (Tr.:BE→NBE)	0.5841	0.806%	0.6983	0.073%
Ours (Tr.:BE→NBE)	0.8111	0.633%	0.8079	0.268%
UniGradICON [40]	0.5326	0.002%	0.8380	0.001%

Fig. 6: **3D registration.** Pixel-based similarity (LNCC) methods andfoundational UniGradICON model fail to capture true correspondences in the "missing anatomy" test scenario (BE→NBE), stretching the brain to the space of skull and neck. Our method reliably performs desired alignment in both scenarios.

As we see in Tab. 2 and Fig. 6, both LNCC [2] networks and foundational UniGradICON [40] fail to capture semantically meaningful correspondences and rely on edges with strong contrast, stretching the gray matter to the skull and neck region, while our method learns semantically meaningful alignment. It significantly outperforms other methods in the test BE→NBE scenario and performs on par in the standard BE→BE test scenario with foundational UniGradICON and an LNCC network that was specifically trained for this scenario.

6   Conclusion

In this work, we presented Diffusion-Guided Image Registration (DGIR), a method inspired by the observation that pre-trained diffusion models trained solely for image generation have the emergent capability of finding semantic correspondences even in medical images. We show that diffusion-guided image registration

is useful in challenging scenarios of missing anatomy (monomodal and multi-modal), where deep semantic understanding is required to find same anatomies and perform accurate registration. A potential future work might be to train a large-scale volumetric diffusion model on multiple diverse datasets and see what emergent properties these models have.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage **54**(3), 2033–2044 (2011)
2. 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)
3. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. ICLR (2021)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: CVPR (2021)
5. Chen, X., Liu, Z., Xie, S., He, K.: Deconstructing denoising diffusion models for self-supervised learning. ICLR (2025)
6. Czolbe, S., Krause, O., Feragen, A.: Semantic similarity metrics for learned image registration. In: Medical Imaging with Deep Learning. pp. 105–118. PMLR (2021)
7. Demir, B., Niethammer, M.: Multimodal image registration guided by few segmentations from one modality. In: MIDL (2024)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
9. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. NeurIPS **34**, 8780–8794 (2021)
10. Fischl, B.: FreeSurfer. Neuroimage **62**(2), 774–781 (2012)
11. Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X.: Deep learning in medical image registration: a review. Physics in Medicine & Biology **65**(20), 20TR01 (2020)
12. Fuest, M., Ma, P., Gui, M., Fischer, J.S., Hu, V.T., Ommer, B.: Diffusion models and representation learning: A survey. arXiv preprint arXiv:2407.00783 (2024)
13. Greer, H., Kwitt, R., Vialard, F.X., Niethammer, M.: ICON: Learning regular maps through inverse consistency. In: ICCV. pp. 3396–3405 (2021)
14. Greer, H., Tian, L., Vialard, F.X., Kwitt, R., Bouix, S., San Jose Estepar, R., Rushmore, R., Niethammer, M.: Inverse consistency by construction for multistep deep registration. In: MICCAI. pp. 688–698. Springer (2023)

15. Haber, E., Modersitzki, J.: Intensity gradient based registration and fusion of multi-modal images. In: MICCAI. pp. 726–733. Springer (2006)
16. 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)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
18. Hoopes, A., Hoffmann, M., Greve, D.N., Fischl, B., Guttag, J., Dalca, A.V.: Learning the effect of registration hyperparameters with hypermorph. MELBA **1** (2022)
19. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR. pp. 9492–9502 (2024)
20. Kim, B., Han, I., Ye, J.C.: Diffusemorph: Unsupervised deformable image registration along continuous trajectory using diffusion models. ECCV (2022)
21. Kögl, F., Reithmeir, A., Sideri-Lampretsa, V., Machado, I., Braren, R., Rueckert, D., Schnabel, J.A., Zimmer, V.A.: General vision encoder features as guidance in medical image registration. In: WBIR MICCAI (2024)
22. Lee, H.Y., Tseng, H.Y., Yang, M.H.: Exploiting diffusion prior for generalizable dense prediction. In: CVPR. pp. 7861–7871 (2024)
23. Luo, G., Dunlap, L., Park, D.H., Holynski, A., Darrell, T.: Diffusion hyperfeatures: Searching through time and space for semantic correspondence. NeurIPS **36** (2023)
24. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience **19**(9), 1498–1507 (2007)
25. Meng, B., Xu, Q., Wang, Z., Cao, X., Huang, Q.: Not all diffusion model activations have been evaluated as discriminative features. NeurIPS (2024)
26. Mukhopadhyay, S., Gwilliam, M., Yamaguchi, Y., Agarwal, V., Padmanabhan, N., Swaminathan, A., Zhou, T., Ohya, J., Shrivastava, A.: Do text-free diffusion models learn discriminative visual representations? In: ECCV. Springer (2024)
27. Nevitt, M., Felson, D., Lester, G.: The osteoarthritis initiative. Protocol for the cohort study **1**, 2 (2006)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. TMLR (2024)
29. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023. pp. 1–11 (2023)
30. Qin, Y., Li, X.: FSDiffReg: Feature-wise and score-wise diffusion-guided unsupervised deformable image registration for cardiac images. In: MICCAI (2023)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
32. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
33. Shi, Y., Xue, C., Liew, J.H., Pan, J., Yan, H., Zhang, W., Tan, V.Y., Bai, S.: DragDiffusion: Harnessing diffusion models for interactive point-based image editing. In: CVPR. pp. 8839–8849 (2024)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
35. Song, X., Xu, X., Yan, P.: General purpose image encoder DINOv2 for medical image registration. MICCAI (2024)

36. Starck, S., Sideri-Lampretsa, V., Kainz, B., Menten, M., Mueller, T., Rueckert, D.: Diff-def: Diffusion-generated deformation fields for conditional atlases. arXiv preprint arXiv:2403.16776 (2024)
37. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al.: UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine **12**(3), e1001779 (2015)
38. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. NeurIPS **36**, 1363–1389 (2023)
39. Tian, J., Aggarwal, L., Colaco, A., Kira, Z., Gonzalez-Franco, M.: Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In: CVPR. pp. 3554–3563 (2024)
40. Tian, L., Greer, H., Kwitt, R., Vialard, F.X., San José Estépar, R., Bouix, S., Rushmore, R., Niethammer, M.: unigradicon: A foundation model for medical image registration. In: MICCAI. pp. 749–760. Springer (2024)
41. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR. pp. 1921–1930 (2023)
42. Tursynbek, N., Niethammer, M.: Unsupervised discovery of 3d hierarchical structure with generative diffusion features. In: MICCAI. pp. 320–330. Springer (2023)
43. Xiang, W., Yang, H., Huang, D., Wang, Y.: Denoising diffusion autoencoders are unified self-supervised learners. In: CVPR. pp. 15802–15812 (2023)
44. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: Fast predictive image registration–a deep learning approach. NeuroImage **158**, 378–396 (2017)
45. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: CVPR. pp. 5729–5739 (2023)
46. Zhuo, Y., Shen, Y.: Diffusereg: Denoising diffusion model for obtaining deformation fields in unsupervised deformable image registration. In: MICCAI (2024)