A Scalable Distributed Framework for Multimodal GigaVoxel Image Registration

Rohit Jena*,† Vedant Zope† Pratik Chaudhari James C. Gee

University of Pennsylvania
Philadelphia, PA 19104, USA
{rjena, zope, pratikac}@seas.upenn.edu, gee@upenn.edu

Abstract

In this work, we propose **FFDP**, a set of IO-aware non-GEMM fused kernels supplemented with a distributed framework for image registration at unprecedented scales. Image registration is an inverse problem fundamental to biomedical and life sciences, but algorithms have not scaled in tandem with image acquisition capabilities. Our framework complements existing model parallelism techniques proposed for large-scale transformer training by optimizing non-GEMM bottlenecks and enabling convolution-aware tensor sharding. We demonstrate unprecedented capabilities by performing multimodal registration of a $100\mu m$ *ex-vivo* human brain MRI volume at native resolution – an inverse problem more than $570\times$ larger than a standard clinical datum in about a minute using only 8 A6000 GPUs. FFDP accelerates existing state-of-the-art optimization and deep learning registration pipelines by upto $6-7\times$ while reducing peak memory consumption by $20-59\%$. Comparative analysis on a $250\mu m$ dataset shows that FFDP can fit upto $64\times$ larger problems than existing SOTA on a single GPU, and highlights both the performance and efficiency gains of FFDP compared to SOTA image registration methods.

1 Introduction

Image registration (also called 'image alignment' or 'image matching') is a non-linear inverse problem ubiquitous in biomedical and life sciences. Given d-dimensional images $F:\Omega\to\mathbb{R}^d$ and $M:\Omega\to\mathbb{R}^d$ defined on domain Ω (usually a compact subset of $\mathbb{R}^d$), image registration seeks to find a coordinate transform $\varphi:\Omega\to\Omega$ that deforms the moving image M to look similar to the fixed image F. Mathematically, we minimize the following objective (Fig. 1):

$$\varphi^* = \arg\min_{\varphi\in G} L(\varphi) \doteq C(F, M\circ\varphi) + R(\varphi) \tag{1}$$

where C is a cost or dissimilarity function, and $\circ$ is the interpolation operator, i.e. $(I\circ g)(x) = I(g(x))$ for all $x\in\Omega$. Popular choices of φ are affine and deformable transforms, i.e. $\varphi(x) = Ax+t$, and $\varphi(x) = x+u(x)$. Modern registration pipelines (Hoffmann et al., 2021; Jena et al., 2024a) consider an affine matching followed by a deformable matching step, resulting in a composite transform $\varphi(x) = Ax+t+u(x)$. u is called the displacement field, modeled as a grid of per-voxel vectors $u(x)\in\mathbb{R}^d$. For an image of size N, the displacement field is a tensor of size dN. We use $[\mathbf{x}]_\Omega$, $A[\mathbf{x}]_\Omega + t$, and $[\mathbf{u}]_\Omega$ to denote the identity grid, grid of affine transformed coordinates, and deformation grid defined on Ω respectively. Common choices of C are mean squared error, Localized Normalized Cross Correlation (Avants et al., 2008a), and Mattes Mutual Information (Mattes et al., 2001). Common choices of R include Sobolev norm of the gradient or warp fields (Beg et al., 2005; Mang et al., 2019; Avants et al., 2008b), total variation, and inverse-consistency (Christensen & Johnson, 2001). To optimize Eq. (1), iterative methods optimize φ^* directly using gradient descent, and deep learning methods learn a deep neural network $\varphi = f_\theta(F, M)$. Image registration establishes a common coordinate system, aligning scans across individuals and atlases (Hering et al., 2022;

*Corresponding author.
†These authors contributed equally.

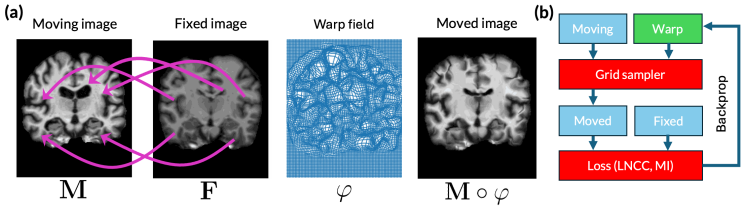


Figure 1: Image Registration Problem. (a): The task is to find a coordinate transform that warps the moving image M to the fixed image F. Individual corresponding points are shown as violet arrows; the per-pixel coordinate transform is shown as a warp field φ, and the transformed image $M \circ \varphi$. **(b)**: A typical registration pipeline - the grid sampler warps the moving image, that is then compared to the fixed image using a loss function. **Green** denotes the optimizable warp, **red** denotes the primary bottlenecks that we optimize in this paper.

Marcus et al., 2007; Murphy et al., 2011). This alignment is a prerequisite for multimodal data fusion, cross-subject comparison, morphometric analysis (Das et al., 2009), and construction of large-scale atlases (Wang et al., 2020b). Establishing such voxelwise correspondence is fundamental for studying anatomical variability, detecting pathological signatures (Ravikumar et al., 2021), and advancing precision medicine (Börner et al., 2022; Jonsson et al., 2022). The saliency and centrality of the task across various biomedical and life science applications has spurred numerous methodological advances in the field, spanning more than three decades of research (Gee et al., 1993; Tian et al., 2024).

Over the past decade, advances in MRI, CT, PET, STPT, and microscopy have enabled ultra-high-resolution imaging, often more than three orders of magnitude larger than macroscopic biomedical domains (Balchandani & Naidich, 2015; Esquivel et al., 2022; Badawi et al., 2019; Gambarotto et al., 2019; Wassie et al., 2019; Kleven et al., 2023; Wang et al., 2020b; Mansour et al., 2025; Kleinfeld et al., 2011). While a typical clinical registration problem involves $\sim$20M parameters, high-resolution ex-vivo human brain scans can require solving up to 11B parameters, far beyond the $\sim$50M-parameter scale at which current registration methods remain reliable. As a result, state-of-the-art deformable image alignment struggles to scale to the resolutions demanded in modern neuroimaging, computational pathology, developmental biology, and connectomics, creating a substantial performance gap. In parallel, innovations in large-scale transformer training such as IO-aware fused operations (Dao et al., 2022; Dao, 2023; Spector et al., 2025) and 5D parallelism for distributing larger-than-memory workloads (Shoeybi et al., 2019; Li et al., 2023; Jacobs et al., 2024; Li et al., 2024; Zhao et al., 2023; Ansel et al., 2024) optimize GEMM-like workflows. However, the fundamental concepts utilized by these methods (IO-awareness, recomputing and aggregating intermediates on shared memory to minimize high bandwidth memory (HBM) storage, identifying partial aggregates across hosts to minimize communication overheads for distributed optimization) are broadly applicable to a wide class of problems of the non-GEMM nature.

In this paper, we apply these concepts to scale image registration algorithms to match parity with the developments in both increasing resolution of image acquisition *and* compute capabilities. To that end, our contributions are twofold. First, we identify key compute and memory bottlenecks in image registration algorithms, and propose novel components that fit problems upto $64\times$ larger than existing algorithms on a single GPU. Second, we propose **F**lash **F**used **D**istributed **P**rimitives (FFDP), a distributed framework to scale registration to an arbitrary number of GPUs, thereby scaling to ultra high-resolution problems. We present a first-of-its-kind demonstration: aligning a $250\mu m$ in-vivo MRI (Lüsebrink et al., 2017) to a $100\mu m$ *ex-vivo* human brain FLASH volume (Edlow et al., 2019) – a multimodal registration problem more than **570**$\times$ larger than a standard clinical datum (Marcus et al., 2007), with over 11.8B transform parameters – completed in *one minute* using only 8 A6000 GPUs. FFDP accelerates existing traditional registration pipelines by upto $7.48\times$ while reducing memory consumption by upto 59%, and deep learning pipelines by upto $6.14\times$ while consuming upto 24% less memory. We highlight the necessity of performing high-resolution registration by comparing our method with various SOTA optimization and deep learning baselines on a $250\mu m$ T1-weighted MRI dataset, showing unprecedented performance and gains in efficiency.

2 Related Work

2.1 Memory Efficient and Large Scale Optimization

Recent years have also witnessed tremendous innovations in large-scale transformer model training. IO-aware implementations typically include individual fused kernels (Dao et al., 2022; Dao, 2023) and domain-specific languages (Spector et al., 2025; PyTorch, 2025) to minimize launch latency and large memory overheads. To distribute larger-than-memory model training workloads across multiple GPUs, 5D parallelism techniques (Shoeybi et al., 2019; Li et al., 2023; Jacobs et al., 2024; Li et al., 2024; Zhao et al., 2023; Ansel et al., 2024) have been proposed. Many of these techniques leverage a divide-and-conquer approach to break down a larger GEMM-like operation like matrix multiplication or attention into smaller sub-problems that can be executed on multiple GPUs and synchronized to compute the final result. To our knowledge, most of these techniques are tailored to transformer-specific architectures and GEMM-like operations (self attention, FeedForward, LayerNorm, etc.) only, and a Model Parallel variant for convolution-aware tensor sharding and synchronization is not available.

2.2 Large Scale Registration in Life Sciences and Biomedical Imaging

***Ex-vivo* neuroimaging.** Neuroanatomical studies often integrate high-resolution ex-vivo MRI, blockface imaging, and histology to bridge the gap between in-vivo imaging and microscopic "gold standard" pathology (Casamitjana et al., 2025; Ravikumar et al., 2024). While large-scale consortia like SEA-AD and HMBA, along with submillimeter whole-brain datasets (Edlow et al., 2019; Lüsebrink et al., 2017), aim to map cellular and molecular organization across species, computational costs often limit analysis to local effects. Specifically, existing tools cannot register these datasets at native resolution due to excessive memory requirements. We overcome this limitation, demonstrating native-resolution registration of whole-brain datasets in one minute using eight A6000 GPUs (see Section 5.2), thereby preserving fine anatomical details typically lost to downsampling.

Large-scale registration in model organisms. Over the past decade, imaging across the life sciences and biomedical domains has progressed from mesoscale surveys to organ- and organism-wide acquisitions at cellular or even subcellular resolution. These span transparent organisms and small animal models (e.g., *C. elegans*, zebrafish, adult *Drosophila*) (Varol et al., 2020; Venkatachalam et al., 2016; Marquart et al., 2017; Gupta et al., 2018; Peng et al., 2011; Brezovec et al., 2024), adult mouse and rat brains imaged at sub-micron resolutions (Gong et al., 2016; Wang et al., 2020a; Kleven et al., 2023) using Light Sheet Fluorescence Microscopy (LSFM) and Serial Two-Photon Microscopy (STPT) imaging. Such modalities routinely generate gigavoxel to teravoxel volumes (Kutten et al., 2016; Nazib et al., 2018). Their scientific utility, however, hinges on the ability to perform registration at the native resolution of acquisition, i.e. aligning specimens (or modalities) in a common coordinate system without sacrificing the fine-scale morphologies including cell bodies, layers, axon bundles, synaptic neighborhoods, etc. that motivate high-resolution acquisition in the first place (Nazib et al., 2018; Goubran et al., 2013).

Across these diverse domains, the unifying requirement demands access to scalable multimodal registration algorithms - a challenge we address in this work. We provide an extended discussion of more related work and the necessity of our approach in Section A.

3 Fused Kernels for Memory Efficient Registration on a Single GPU

Bottlenecks of a deformable image registration pipeline Our primary objective is to identify compute and memory bottlenecks in *large-scale* image matching tasks. In identifying these bottlenecks, training-free optimization methods are better suited than deep networks since the latter has a much larger activation memory footprint, which forms the primary memory bottleneck (Tazi et al., 2024). For instance, for a $250\mu m$ image pair, a standard deep learning method (Hoffmann et al., 2021) generates an activation map of size 27GB only after the first layer. Extrapolating memory usage for clinical data, existing deep networks will require upto 1.2TB of GPU memory at inference to process these image volumes at native resolution. In contrast, a training-free optimizer can fit this problem in less than 45GB of GPU memory. We use FireANTs (Jena et al., 2024a) as our base framework to identify compute and memory bottlenecks in a typical image registration problem. We analyze

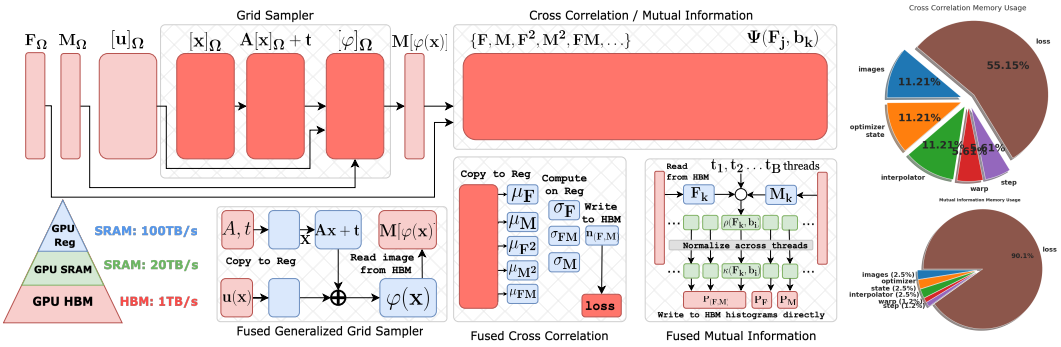


Figure 2: Left: FFDP uses fused kernels to eliminate intermediate HBM memory usage (in **dark red**) for memory-bound workhorse operations (`grid_sampler`, LNCC, MI) for large-scale image registration. For `grid_sampler` and LNCC, additional intermediate per-pixel variables (warp coordinates, patchwise statistics) are computed per-pixel in registers (blue). For MI, the Parzen Windowing and histogram aggregation is performed using shared memory (green), avoiding large HBM overheads. **Right**: Pie charts show the breakdown of memory overheads for storing the image, grid, optimizer state, and intermediate variables for MI and LNCC losses.

the flamegraph of a typical clinical MRI registration task from the *OASIS* brain dataset (Marcus et al., 2007) in Fig. 20. We identify three key memory bottlenecks in image matching pipelines (1) deformable interpolation and warp composition (2) cross-correlation loss, and (3) mutual information loss (see Fig. 2(right)). [1] We first propose efficient designs to fit larger problems on a single GPU, and then extend the framework to distributed registration.

3.1 Composite Implicit Grid Sampler

A fundamental operation used in image registration is the *grid sampler*. This operator allows us to warp an image M using a deformation field $\varphi : \Omega \to \Omega$ and computes the image $M' : M'(x) = M(\varphi(x))$. Virtually every image registration pipeline uses this operation to warp the moving image using an affine, deformable, or composite transform. For affine and composite transforms, the operator initializes a regular grid $[\mathbf{x}]_\Omega$, a grid of size $3N$. The affine grid $A[\mathbf{x}]_\Omega + t$ is another grid of size $3N$. If a deformable grid $[\mathbf{u}]_\Omega$ is optimized, then a third grid $A[\mathbf{x}]_\Omega + t + [\mathbf{u}]_\Omega$ is materialized, costing a total of $9N$ overhead for an image of size N. To consolidate these memory overheads, we propose a composite implicit grid sampler. This is a fused CUDA kernel that performs the following operation:

$$\texttt{fused\_grid\_sampler}(I; A, t, [\mathbf{u}], S, x_{\text{bounds}})(x) = I(Ax + t + Su(x))$$

where $A, S \in GL(d, \mathbb{R})$ are affine matrices, t is a translation vector, $[\mathbf{u}]$ is the deformation grid, and x_{bounds} are the bounds of the (implicit) identity grid $[\mathbf{x}]_\Omega$. There are three benefits of this approach. First, the kernel avoids materializing any additional grids in HBM, reducing the memory overhead of the kernel from $O(n)$ to $O(1)$ with no loss in runtime or accuracy. Second, when the warp $[\mathbf{u}]_\Omega$ is sharded across hosts in a distributed setting, the identity grid $[\mathbf{x}]_\Omega$ needs to be sharded correctly too. Since the identity grid is implicitly defined by its bounds $x_{\text{bounds}} = (x_{\min}, x_{\max}) \in \mathbb{R}^{2d}$, our implementation can be easily used in a distributed optimization setting without instantiating partial shards $[\mathbf{x}]_{\Omega_h}$. Finally, the matrix S is used to rescale the deformation field to sample from the coordinates of the sharded images I_h which lie on the grid Ω_h instead of Ω (see Section I.2) without initializing additional memory. The backward pass is very similar to the existing PyTorch implementation, with the exception of the gradient of the affine matrix. We discuss the derivation and pseudocode of the forward and backward pass in the Section H.

3.2 Implicit Parzen Windowing for Mutual Information

Mattes Mutual Information (MI) is one of the most commonly used loss functions for *multimodal* image matching (Chen et al., 2022; Avants et al., 2009; Mattes et al., 2001). For random variables X and Y, MI is the KL divergence between the joint distribution $P(X, Y)$ and product of marginal

[1] A GPU's memory hierarchy spans multiple tiers: registers (per-thread, single-cycle), shared memory/L1 cache (on-chip, tens of KB, low latency within a block), L2 cache (MBs, shared across SMs, moderate latency), and global memory (**HBM**). Our work focuses on reducing HBM usage for key non-GEMM operations used in image registration, by maximizing register and shared memory usage while minimizing global memory traffic.

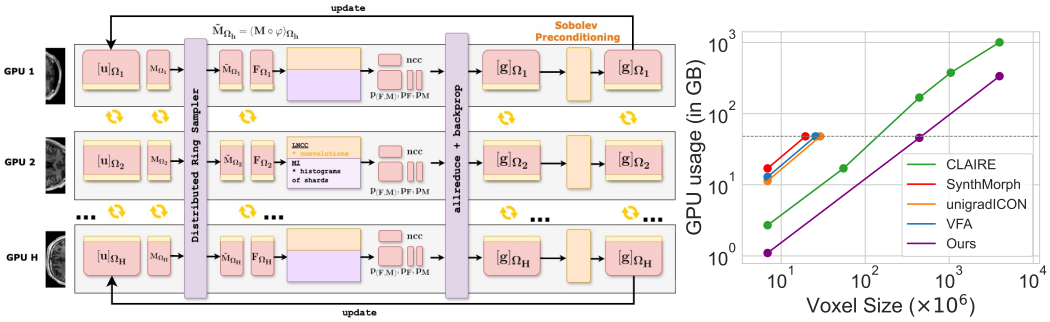


Figure 3: Left: Overview of our distributed framework. GridParallel (GP) shards the fixed and moving images (F, M) and the warp field $[\mathbf{u}]$ across multiple GPUs. Yellow blocks and arrows denote synchronized halo boundaries between GPUs, enabling smoothing on images and warp fields without an allgather. The ring sampler (violet) computes interpolated image shards on the fly, avoiding materialization of the full moving image. We then compute losses (MSE, LNCC, MI), compute gradients w.r.t. each warp shard, apply Sobolev regularization with GP, and update shards by gradient descent. **Right**: Scaling efficiency compared to deep methods and CLAIRE (Mang et al., 2019), a distributed registration method. Most SOTA deep learning baselines require orders-of-magnitude more memory for the same problem size and scalability is limited to a single GPU (dotted line). Our framework scales to arbitrarily large problem sizes while using about $5\times$ less memory than CLAIRE.

distributions $P(X)P(Y)$ of the intensities of the two images. For image matching, X and Y are the pixel intensities for the images I, J. The distributions are estimated using a kernel density estimator:

$$P_I(v) = \frac{1}{N}\sum_k \kappa(v - I_k), \quad P_{(I,J)}(v, w) = \frac{1}{N}\sum_k \kappa(v - I_k)\kappa(w - J_k) \tag{2}$$

where κ is a kernel function of choice. Common choices of κ are the Gaussian (Guo, 2019) and 3rd order B-Spline kernels (Thévenaz & Unser, 2000). To empirically compute the KL divergence, the distributions Eq. (2) are discretized over B equally spaced bins on the domain of $u \in I, v \in J$. However, to compute the joint histogram of size B^2, this method requires materializing the entire Parzen Block $\Psi_I(j, k) = \kappa(b_j - I_k)$ of size $2k_P BN$, where k_P is a kernel-dependent constant. Since $N >> B$ (B is typically chosen to be 32), this operation becomes a significant memory bottleneck for large N. For instance, a typical clinical image volume ($N \approx 30$MB) with 32 bins will consume **7.5GB** of HBM - a significantly huge cost that grows much faster for larger problems.

Our efficient implementation leverages the fact that B is small to avoid materializing the tensors $\Psi_I, \Psi_J \in \mathbb{R}^{B\times N}$ altogether and use high-throughput shared memory to compute and accumulate the histogram entries and partial gradients for each image pixel. We provide the detailed derivation in Section G. This leads to an efficient implementation that consumes $O(1)$ additional HBM instead of $O(N)$ (holding B constant). This leads to upto **98%** lesser HBM usage for images considered in our experiments, and an asymptotic 100% reduction in HBM usage for large images (Fig. 7(top-right)).

3.3 Efficient Implicit Fused Cross-Correlation

Local Normalized Cross-Correlation (LNCC) is used ubiquitously in signal and image processing as a similarity metric. In deformable image registration, it is used as a robust similarity function to compare anatomical similarities (Chen et al., 2022; Hoffmann et al., 2021; Avants et al., 2008b; Wu et al., 2024). Most LNCC implementations are memory-bound due to the large number of intermediate variables. Our analysis in Section F shows that the computational graph adds $16\times$ HBM overhead, and upto another $16\times$ HBM overhead for computing gradients with respect to all intermediates.

To avoid these huge memory overheads, we fuse all the intermediate computation in a fused kernel. Our fused forward pass requires only $5\times$ memory for storing all intermediates (I, J, I^2, J^2, IJ convolved with matrix w). In Section F we analytically derive the gradient and show that the input gradients can be computed by modifying the saved intermediates *in-place*. This leads upto a **76.5% reduction** in memory (see Table 3) and outperforms even `torch.compile` implementations.

4 Extending Image Registration to Multiple GPUs

Our composite implicit grid sampler and improved loss functions allows optimizing problems with image sizes that are upto two magnitudes larger than other baselines on a single A6000 GPU (Fig. 5a).

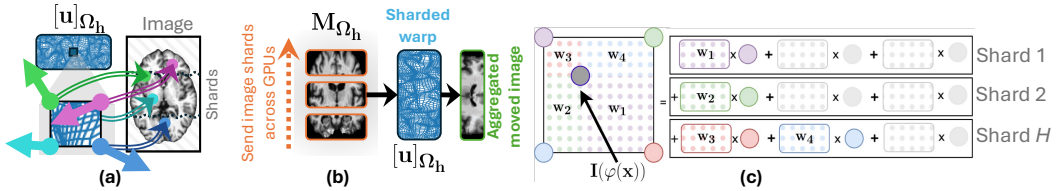


Figure 4: **(a)** Neighboring coordinates in the warp field may refer to pixel locations on arbitrary image shards due to the deformable nature of the warp field, making distributed interpolation non-trivial. **(b)** Ring Sampler interleaves fetching of image shards and aggregating the partial sums of interpolated values, avoiding a memory-expensive allgather. **(c)** Bilinear Interpolation is decomposed into partial sums over image shards, which are accumulated with a ring topology communication, similar to Liu et al. (2024b).

However, many applications using mesoscopic and microscopic data require registration of images that do not fit on a single GPU. Inspired by distributed frameworks for LLM training (Shoeybi et al., 2019; Rajbhandari et al., 2020) and initial work on distributed image registration (Mang et al., 2019), we propose a distributed framework that allows sharding large images across multiple GPUs to efficiently scale to arbitrarily large problem sizes with any similarity loss function.

Distributed Setting. For distributed registration with H hosts or GPUs, we partition the domain $P(\Omega) = \{\Omega_1, \Omega_2, \dots \Omega_H\}$ such that $|\Omega_i| = N/H$, $\Omega_i \cap \Omega_j = \phi \quad \forall i \neq j$ and $\cup_i \Omega_i = \Omega$. We use $[\mathbf{x}]_{\Omega_h}$, $A[\mathbf{x}]_{\Omega_h} + t$, and $[\mathbf{u}]_{\Omega_h}$ to denote the sharded tensors defined on domain Ω_h.

4.1 Grid Parallel for Boundary-Synchronized Image Sharding

Techniques like Tensor/Sequence/Expert/Context Parallel have been tremendously successful in distributed optimization by sharding large models and sequences across multiple GPUs (Shoeybi et al., 2019; Li et al., 2023; Liu et al., 2024b;a). However, these techniques work for transformer-like architectures and input sequences where the model parameters and activations do not require boundary synchronization. In contrast, image registration contains operations that require boundary synchronization between image and grid shards to perform mathematically correct convolutions. Examples of such operations include convolutions for calculating LNCC, total variation loss, Sobolev norm of the gradient and warp fields (Mang et al., 2019; Avants et al., 2008b; Beg et al., 2005).

To enable these functionalities and complement existing parallelism techniques, we propose '*Grid Parallel*' (GP) as an abstraction on a tensor. GP shards a tensor across hosts, stores the sharded dimension and bounds as metadata, and provides synchronization operations to augment the tensor with sufficient boundary padding from neighboring shards prior to performing a convolution operation. GP allows us to partition the fixed images, $[\mathbf{u}]$, and the optimizer state $[\mathbf{m_1}], [\mathbf{m_2}]$ – essentially sharding the entire problem across H hosts while allowing the user to apply convolutional operations seamlessly. We compare the performance of GP with naive DTensor sharding in Section D.

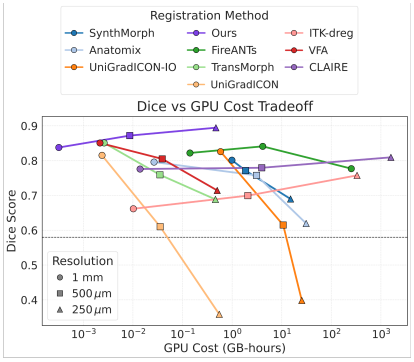
4.2 Distributed Ring Sampler

Despite the sharding in GP, the moving image M cannot be sharded across GPUs due to the random-access nature of the `grid_sample` operation applied on M. In general, the warp vector $\varphi(x)$ residing on GPU i can point to coordinates that reside on the sharded image on GPU j for any $j \neq i$. Even for neighboring coordinates $x_s, x_u \in [\mathbf{x}]_i$, the coordinates $\varphi(x_s)$ and $\varphi(x_u)$ can point to different shards $j_1 \neq j_2 \neq i$. This is illustrated in Fig. 4(a). Keeping the entire moving image in memory limits the maximum problem size to $N \leq V$, where V is the memory per GPU, regardless of the number of hosts H. However, we want the maximum problem size to scale with H. Therefore, we propose a distributed `grid_sampler` that allows us to *correctly* interpolate the moving image with sharded images scattered across multiple hosts without performing an `allgather` operation on the moving image.

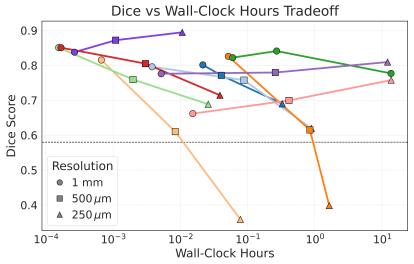
Our approach leverages the key observation that (bi/tri)linear interpolation can be decomposed as an aggregate of partial sums of interpolated values on individual image shards. Fig. 4(b) illustrates this example. These individual image shards are sent across hosts in a ring topology, similar to Liu et al. (2024b), and the partial sum is aggregated inplace. This operation only incurs an additional N/H HBM overhead for fetching the sharded image from other hosts, scaling efficiently to arbitrary large

(a) Performance comparison across methods and resolutions.

Resolution	Method	AvgDice Score ↑	InvDice Score ↑	AvgHD90^{cum} (mm) ↓
1 mm	Baseline	0.579 ± 0.055	0.141 ± 0.142	1.587 ± 0.908
	Anatomix	0.796 ± 0.035	0.386 ± 0.138	0.468 ± 0.137
	CLAIRE	0.776 ± 0.044	0.344 ± 0.120	0.554 ± 0.150
	FireANTs	0.822 ± 0.032	0.435 ± 0.147	0.393 ± 0.126
	ITK-dreg	0.662 ± 0.055	0.199 ± 0.125	1.002 ± 0.277
	SynthMorph	0.801 ± 0.022	0.378 ± 0.133	0.455 ± 0.098
	TransMorph	0.851 ± 0.016	0.468 ± 0.161	0.310 ± 0.064
	UniGradICON (IO)	0.826 ± 0.022	0.391 ± 0.155	0.384 ± 0.095
	UniGradICON	0.815 ± 0.026	0.393 ± 0.156	0.419 ± 0.113
	VFA	0.851 ± 0.023	0.494 ± 0.169	0.323 ± 0.096
	Ours	0.838 ± 0.028	0.436 ± 0.148	0.341 ± 0.109
500 µm	Baseline	0.580 ± 0.055	0.138 ± 0.143	1.357 ± 0.326
	Anatomix†	0.758 ± 0.040	0.325 ± 0.159	0.619 ± 0.169
	CLAIRE	0.779 ± 0.051	0.275 ± 0.210	0.570 ± 0.211
	FireANTs	0.841 ± 0.033	0.489 ± 0.163	0.340 ± 0.127
	ITK-dreg	0.699 ± 0.056	0.240 ± 0.130	0.834 ± 0.254
	SynthMorph†	0.771 ± 0.035	0.337 ± 0.133	0.557 ± 0.144
	TransMorph†	0.759 ± 0.028	0.300 ± 0.175	0.624 ± 0.127
	UniGradICON†	0.610 ± 0.044	0.133 ± 0.122	1.231 ± 0.262
	UniGradICON (IO)†	0.615 ± 0.047	0.149 ± 0.136	1.527 ± 1.495
	VFA†	0.805 ± 0.044	0.419 ± 0.181	0.462 ± 0.163
	Ours	0.872 ± 0.028	0.528 ± 0.180	0.258 ± 0.099
250 µm	Baseline	0.580 ± 0.055	0.136 ± 0.141	1.409 ± 0.322
	Anatomix†	0.620 ± 0.031	0.161 ± 0.115	1.179 ± 0.190
	CLAIRE	0.809 ± 0.054	0.378 ± 0.133	0.570 ± 0.211
	FireANTs†	0.777 ± 0.064	0.341 ± 0.199	0.629 ± 0.295
	ITK-dreg	0.758 ± 0.048	0.299 ± 0.125	0.613 ± 0.191
	SynthMorph†	0.690 ± 0.052	0.243 ± 0.164	0.882 ± 0.239
	TransMorph†	0.689 ± 0.044	0.191 ± 0.132	0.973 ± 0.245
	UniGradICON (IO)†	0.398 ± 0.062	0.063 ± 0.071	3.491 ± 3.198
	UniGradICON†	0.359 ± 0.044	0.045 ± 0.056	2.992 ± 0.670
	VFA†	0.714 ± 0.066	0.281 ± 0.216	0.821 ± 0.300
	Ours	0.895 ± 0.029	0.597 ± 0.204	0.216 ± 0.098



(b) Accuracy vs. GPU Compute Cost.



(c) Accuracy vs. Wall-clock Time.

Figure 5: Registration performance on Faux-OASIS dataset at 1 mm, 500 µm, and 250 µm (native 250 µm); mean ± std over pairs. ↑ higher is better; ↓ lower is better. HD90 values are reported using our cumulative definition (see Sec. K.2). (Green)/ (Yellow) = best/second; †= patch-based

problem sizes for sufficiently large H. The detailed derivation and correctness of this operation is shown in Section I.

4.3 Distributed Loss Functions

Since the moved image and fixed image are sharded cross H hosts, the loss function must take this into account to compute the loss function correctly.

Mean Squared Error (MSE). Since MSE is a per-pixel loss, we compute the individual MSE on host h and perform an `allreduce` operation.

Localized Normalized Cross Correlation (LNCC). The LNCC computes per-pixel patch similarities for each pixel, using a convolution over its neighbors. For sharded images, the patch statistics at the boundary requires a boundary synchronization with its neighboring shards which is provided by our GP implementation. After computing the LNCC for all pixels in each shard, we perform another `allreduce` to compute the LNCC over the entire image.

Mutual Information (MI). The MI loss computes the joint histograms $p_{(I,J)}(x, y)$ and marginals $p_I(x), p_J(y)$. However, these distributions are partial aggregates from the sharded images on each GPU. Eq. (2) can be rewritten as $p_I(v) = \sum_h \frac{N_h}{N} \left(\frac{1}{N_h} \sum_{k \in \Omega_h} \kappa(v - I_k) \right), p_{IJ}(v, w) = \sum_h \frac{N_h}{N} \left(\frac{1}{N_h} \sum_{k \in \Omega_h} \kappa(v - I_k) \kappa(w - J_k) \right)$, where the red terms correspond to the per-host histogram computation. Performing an `allreduce` to compute the weighted average of these histograms (with weights N_h/N) results in a valid and correct joint and marginal distributions over all hosts. This also leads to only a $B^2 + 2B$ communication overhead regardless of N, making a distributed implementation highly practical.

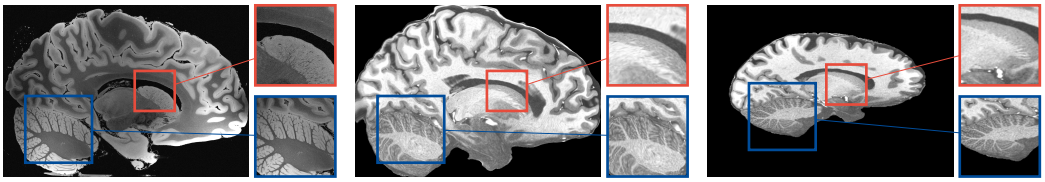


Figure 6: Qualitative comparison on registration of 100μm ex-vivo brain MRI (T1 $\rightarrow$ FLASH) image. Fine details like cerebellar white matter are not visible at macroscopic scales, but are aligned at 100μm. Fixed image is of size $1760 \times 1760 \times 1278$. Best viewed zoomed in. More results in Fig. 11.

5 Experiments

Our primary goals are to (a) accelerate both optimization and neural network based registration workflows, and (b) solve significantly larger image registration problems. We show the efficacy of our method by accelerating existing registration workflows on standard clinical data. This is followed by optimizing a multimodal registration task with more than 11.8B optimizable parameters, an unprecedented result in large-scale registration. We compare the performance and computational efficiency of our method with various state-of-the-art baselines on a simulated 250μm ex-vivo brain MRI dataset, followed by ablations on various components of our framework.

Baselines. To accelerate existing registration workflows, we compare against TransMorph (Chen et al., 2022) and FireANTs (Jena et al., 2024a), which are state-of-the-art deep learning and optimization based registration frameworks respectively. In addition, we perform comparative evaluation with two methods explicitly designed for large-scale registration: ITK-DReg (itk) (CPU-based) and CLAIRE (Mang et al., 2019) (multi-GPU), and several SOTA learning-based approaches for clinical data - SynthMorph (Hoffmann et al., 2021), Vector-Field Attention (Liu et al., 2024c), unigradICON (Tian et al., 2024) (with/without instance optimization), anatomix+ConvexAdam (Dey et al., 2025).

5.1 Accelerating existing registration workflows and ablations

For deep networks, we train `TransMorph-large` under three loss configurations: (a) LNCC+Dice, (b) MI+Dice, and (c) LNCC+scaling-and-squaring (Ashburner, 2007) +Dice. For each configuration shown in Table 1, we either use the vanilla PyTorch implementation (Baseline) or our kernels (Ours). For classical optimization, we benchmark runtime and memory against multiple LNCC backends (FireANTs, VoxelMorph/TransMorph, Fast LNCC, `torch.compile`, and Ours) and MI backends (PyTorch and Ours with and without `torch.compile`). Tables 1 and 4 and Fig. 12 show that during network training our kernels converge 6.1$\times$ faster with LNCC while using 16.5% less memory, and reduce MI memory usage by 24.7%. Despite being designed for very large images, the runtime and memory benefits are significant for clinical-scale data (i.e., 30MB for OASIS). Optimization frameworks see larger gains: FireANTs achieves up to 95.2% memory savings and 2.6$\times$ speedup with MI, and a 7.5$\times$ speedup over FastLNCC (Jia et al., 2025) (and 2.9$\times$ over FireANTs' LNCC backend which applies separable convolutions on FastLNCC), with 44-59% lower memory usage overall.

5.2 Registration to a 100 micron ex-vivo brain MRI volume

To showcase the efficacy of our method on real large scale images, we register a 250μm in-vivo MRI image (Lüsebrink et al., 2017) to a 100 μm *ex-vivo* FLASH human brain volume (Edlow et al., 2019). This represents an inverse problem with more than 11.2B optimizable parameters (compared to $\sim$20M for clinical datasets), or 44.8GB of GPU memory. The entire problem does not fit on most GPUs, necessitating distributed multimodal registration. We optimize a composite transform - affine followed by a diffeomorphic mapping; details can be found in Section E.1. Multimodal deformable registration took $\sim$58 seconds on 8 NVIDIA A6000 GPUs, which is unprecedented at this resolution. Fig. 6 shows qualitative results, highlighting the ability to register highly detailed structures such as cerebellar white matter; these structures are not visible at macroscopic scales. The resultant advantages of performing registration at this scale can allow researchers to characterize the

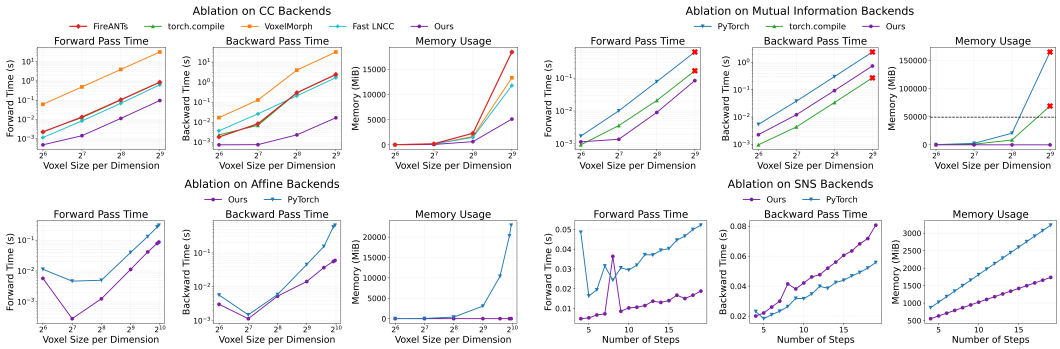


Figure 7: Ablations on key workhorse operations: LNCC, MI, `grid_sampler`, and scaling-and-squaring operations. Our fused kernels consume significantly less HBM and runtime.

neuroanatomy at microscopic resolutions and allow morphometric analysis of cortical layers and subcortical nuclei among other structures.

Registration accuracy in these studies is measured using privately annotated fiducial markers, hindering reproducibility and comparability of methodological advances. Due to lack of scalable frameworks, most high-resolution studies simply run ANTs at a significantly downsampled resolution (Kleven et al., 2023; Mansour et al., 2025; Wang et al., 2020b; Kronman et al., 2024; Bogovic et al., 2020; Edlow et al., 2019) and upsample the warp field to the native resolution.

5.3 Comparative Analysis on a Simulated ex-vivo Brain MRI Dataset

The faux-OASIS dataset To compare registration performance at high resolutions and leverage existing methods as baselines, we synthesize the *faux-OASIS* dataset, which mimics the anatomical distribution of an MRI dataset at $250\mu m$ isotropic resolution (more details in Section K). At $250\mu m$, the deformation field has 1.32B degrees of freedom per image pair, compared to $\sim$20M for OASIS.

Table 1: Accelerating TransMorph (**Top**) and FireANTs (**Bottom**) training with various computation backends.

Variant	Loss	Diffeomorphic	Training Time (h)	GPU Mem (GB)	Val DSC
Baseline	LNCC	✗	171.20	20.01	86.74
Ours	LNCC	✗	27.84	16.95	87.23
Baseline	LNCC	✓	171.42	21.28	86.55
Ours	LNCC	✓	27.93	17.34	87.09
Baseline	MI	✗	26.09	22.34	86.74
Ours	MI	✗	24.94	16.80	86.80

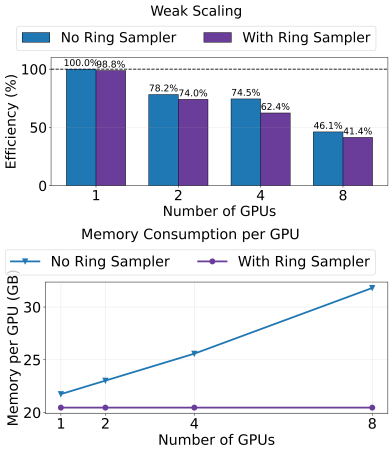
Loss	Backend	Dice Score ↑	Runtime (s) ↓	Memory (MB) ↓
LNCC	FireANTs	78.81 ± 3.87	1.44 ± 0.08	1044.5 ± 0.0
LNCC	FastLNCC	76.96 ± 3.60	3.76 ± 0.16	1026.3 ± 0.0
LNCC	VXM/TM	76.96 ± 3.60	57.08 ± 2.45	1418.5 ± 0.0
LNCC	`torch.compile`	69.35 ± 4.09	0.82 ± 0.04	860.7 ± 0.0
LNCC	Ours	78.67 ± 3.04	0.50 ± 0.01	577.5 ± 0.0
MI	PyTorch	75.88 ± 3.45	7.51 ± 0.37	12206.3 ± 0.0
MI	`torch.compile`	75.88 ± 3.45	1.05 ± 0.05	3865.5 ± 0.0
MI	Ours	75.88 ± 3.44	2.90 ± 0.16	577.5 ± 0.0
MI	`torch.compile`+Ours	75.93 ± 3.47	2.95 ± 0.16	657.3 ± 0.0

Baselines and evaluation. All methods (including CLAIRE and FireANTs without FFDP) run out of memory at $250\mu m$ resolution. We proposed two modifications to deep learning based methods to enable them to work on this dataset: (a) inspired by several high-resolution studies (Wang et al., 2020b; Mansour et al., 2025; Edlow et al., 2019), we register the images at a downsampled resolution, and then upsample the deformation field (b) inspired by several histology registration methods (Wodzinski et al., 2024; Lotz et al., 2015; Liang et al., 2021), we perform patchwise registration and mosaicing of the final deformation. We compare the methods at three resolutions: 1mm, $500\mu m$, and $250\mu m$. At 1mm, the full image fits within a patch, providing a baseline reference comparable to reported OASIS performance. At higher resolutions, patches are defined by each method's default input size with stride equal to 50% of the patch size. FireANTs augmented with FFDP is denoted as *Ours*. We report Dice, inverse-weighted Dice (InvDice; Mang et al. (2019)), and average Haussdorf distance capped at 90 percentile (AvgHD90). To compare efficiency, we measure both wall-clock time and GPU-hours.

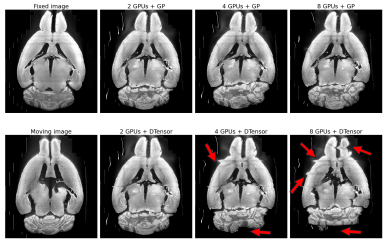
Results. Fig. 5a summarizes performance metrics. At 1mm, most methods achieve performance consistent with their reported performance on OASIS, including VFA and TransMorph which were trained on the OASIS dataset with label supervision. At higher resolutions, nearly all methods degrade, especially for InvDice and HD90, which emphasize alignment of fine structures. In contrast, our method improves in accuracy: at $250\mu m$, we improve Dice by 18.1 points, InvDice by 31.6 points, and reduce AvgHD90 by 62.1%. The correlation between resolution and performance is also

observed in (Mang et al., 2019; Mang & Ruthotto, 2017; Nazib et al., 2018); in addition we verify that patch-based methods *degrade* in performance at higher resolutions.

Figure 8: Scaling and GP ablations.



(a) Weak scaling and Per-GPU memory consumption of FFDP.



(b) Qualitative ablation of GP synchronization in FFDP on the fMOST mouse brain dataset (Tustison et al., 2024). Red arrows highlight regions affected by incorrect boundary effects due to no GP. See Fig. 10 for more examples.

This degradation among patchwise methods is expected; histology-style pipelines typically register consecutive slides with small deformations after affine alignment. At high resolution, patching reduces anatomical context and the patches become progressively more out-of-distribution (see Fig. 19). Patchwise or downsampling strategies are therefore insufficient for ultra-high resolution large-scale registration, and existing deep methods cannot be repurposed to work at higher resolutions efficiently. Accuracy-efficiency tradeoffs in Figs. 5b and 5c show that our method is Pareto-efficient compared to all other methods (CPU, deep learning, and distributed GPU methods), requiring up to $500\times$ fewer GPU-hours compared to alternatives at $250\mu m$.

5.4 Ablation Studies

We ablate on the efficiency of various workhorse operations used in image registration in Fig. 7 and Table 3. We compare our implementations to community-standard PyTorch implementation (Jia et al., 2025; Chen et al., 2022) and `torch.compile` versions. For grid sampler and MI kernels, our kernels have $O(1)$ extra HBM overhead instead of $O(N)$ in the PyTorch implementation. For LNCC, our implementation achieves an average speedup in the forward pass by $5.22\times$ and $56.98\times$ in the backward pass. Our `grid_sampler` also leads to an efficient scaling-and-squaring operation, commonly used in deep learning registration pipelines (Chen et al., 2022), with a memory reduction of 50% compared to the baseline implementation. **Scalability Analysis.** We test the weak scaling of our distributed framework by registering synthetic images with increasing voxel sizes. For H GPUs, we instantiate an image pair of size $700 \times 700 \times 700H$ and shard the images, warp, and optimizer state across H GPUs. Fig. 8a shows weak scaling of FFDP with and without ring sampler. Without the ring sampler, the `grid_sample` operation requires storing the moving image of size $700 \times 700 \times 700H$ on each GPU, leading to peak HBM memory increasing linearly with H. This implies the framework would not scale to arbitrarily large problem sizes, regardless of cluster size H. Peak Memory consumption is independent of H with the Ring Sampler, and scaling efficiency is only minimally affected. **Ablation on GP.** We ablate the effect of GP by replacing it with DTensor sharding (no boundary sync). Figs. 8b, 9 and 10 show that incorrect boundary synchronization leads to undesirable artifacts in the moved images, and reduces labelmap overlap.

6 Conclusion

We propose a novel distributed framework for arbitrarily large image registration problems. Our work identifies and proposes IO-aware and distributed-friendly implementations of workhorse operations in image registration algorithms, enabling registration of images at arbitrarily large resolutions on a single GPU. Our fused primitives demonstrate compelling results in both improving existing registration pipelines and scaling to arbitrarily large, multimodal problems pertinent in modern life science applications, that were previously infeasible without approximations. FFDP shows unprecedented registration capabilities that will enable researchers to leverage and effectively work with large-scale image volumes and unearth new insights leveraging the large resolution images.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) under grants EB031722 and NS135568.

References

Allen brain atlas. URL https://atlas.brain-map.org/.

Itk-dreg: A framework for distributed, large-scale image registration. URL https://itk-dreg.readthedocs.io/en/latest/.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Jesper LR Andersson, Mark Jenkinson, Stephen Smith, et al. Non-linear registration, aka spatial normalisation fmrib technical report tr07ja2. *FMRIB Analysis Group of the University of Oxford*, 2 (1):1–22, 2007.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pp. 929–947, 2024.

ANTsX. Antsx: Advanced normalization tools (ants). URL https://github.com/ANTsX/ANTs. GitHub repository.

John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

John Ashburner. Spm: a history. *Neuroimage*, 62(2):791–800, 2012.

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. *Medical Image Analysis*, 12(1):26–41, February 2008a. ISSN 1361-8423. doi: 10.1016/j.media.2007.06.004.

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. *Medical Image Analysis*, 12(1):26–41, February 2008b. ISSN 1361-8415. doi: 10.1016/j.media.2007.06.004. URL https://www.sciencedirect.com/science/article/pii/S1361841507000606.

Brian B. Avants, P. Thomas Schoenemann, and James C. Gee. Lagrangian frame diffeomorphic image registration: Morphometric comparison of human and chimpanzee cortex. *Medical Image Analysis*, 10(3):397–412, June 2006. ISSN 13618415. doi: 10.1016/j.media.2005.03.005. URL https://linkinghub.elsevier.com/retrieve/pii/S1361841505000411.

Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). *Insight j*, 2 (365):1–35, 2009.

Ramsey D Badawi, Hongcheng Shi, Pengcheng Hu, Shuguang Chen, Tianyi Xu, Patricia M Price, Yu Ding, Benjamin A Spencer, Lorenzo Nardo, Weiping Liu, et al. First human imaging studies with the explorer total-body pet scanner. *Journal of Nuclear Medicine*, 60(3):299–303, 2019.

Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, August 2019. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2019.2897538. URL http://arxiv.org/abs/1809.05231. arXiv:1809.05231 [cs].

P Balchandani and TP Naidich. Ultra-high-field mr neuroimaging. *American Journal of Neuroradiology*, 36(7):1204–1215, 2015.

Erin S Beck, Pascal Sati, Varun Sethi, Tobias Kober, Blake Dewey, Pavan Bhargava, Govind Nair, Irene C Cortese, and Daniel Salo Reich. Improved visualization of cortical lesions in multiple sclerosis using 7t mp2rage. *American Journal of Neuroradiology*, 39(3):459–466, 2018.

M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61:139–157, 2005.

Ganesh Bikshandi and Jay Shah. A case study in cuda kernel fusion: Implementing flashattention-2 on nvidia hopper architecture using the cutlass library. *arXiv preprint arXiv:2312.11918*, 2023.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023.

John A Bogovic, Hideo Otsuna, Larissa Heinrich, Masayoshi Ito, Jennifer Jeter, Geoffrey Meissner, Aljoscha Nern, Jennifer Colonell, Oz Malkesman, Kei Ito, et al. An unbiased template of the drosophila brain and ventral nerve cord. *Plos one*, 15(12):e0236495, 2020.

Katy Börner, Andreas Bueckle, Bruce W Herr, Leonard E Cross, Ellen M Quardokus, Elizabeth G Record, Yingnan Ju, Jonathan C Silverstein, Kristen M Browne, Sanjay Jain, et al. Tissue registration and exploration user interfaces in support of a human reference atlas. *Communications Biology*, 5(1):1369, 2022.

Bella E Brezovec, Andrew B Berger, Yukun A Hao, Feng Chen, Shaul Druckmann, and Thomas R Clandinin. Mapping the neural dynamics of locomotion across the drosophila brain. *Current Biology*, 34(4):710–726, 2024.

Xiaohuan Cao, Jianhua Yang, Jun Zhang, Dong Nie, Minjeong Kim, Qian Wang, and Dinggang Shen. Deformable image registration based on similarity-steered cnn regression. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pp. 300–308. Springer, 2017.

Adrià Casamitjana, Matteo Mancini, Eleanor Robinson, Loïc Peter, Roberto Annunziata, Juri Althonayan, Shauna Crampsie, Emily Blackburn, Benjamin Billot, Alessia Atzeni, et al. A probabilistic histological atlas of the human brain for mri segmentation. *Nature*, pp. 1–8, 2025.

Junyu Chen, Eric C. Frey, Yufan He, William P. Segars, Ye Li, and Yong Du. TransMorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, November 2022. ISSN 13618415. doi: 10.1016/j.media.2022.102615. URL http://arxiv.org/abs/2111.10480. arXiv:2111.10480 [cs, eess].

Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. TVM: An automated End-to-End optimizing compiler for deep learning. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pp. 578–594, Carlsbad, CA, October 2018. USENIX Association. ISBN 978-1-939133-08-3. URL https://www.usenix.org/conference/osdi18/presentation/chen.

Gary E Christensen and Hans J Johnson. Consistent image registration. *IEEE transactions on medical imaging*, 20(7):568–582, 2001.

Gilberto Corso, Gabriel MF Ferreira, and Thomas M Lewinsohn. Mutual information as a general measure of structure in interaction networks. *Entropy*, 22(5):528, 2020.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359, 2022.

Sandhitsu R Das, Brian B Avants, Murray Grossman, and James C Gee. Registration based cortical thickness measurement. *Neuroimage*, 45(3):867–879, 2009.

Chris Davis and A Murat Maga. Image registration and template based annotation of great ape skulls. In *American Journal of Physical Anthropology*, volume 165, pp. 60–61. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2018.

Bob D De Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.

Neel Dey, Benjamin Billot, Hallee E. Wong, Clinton Wang, Mengwei Ren, Ellen Grant, Adrian V Dalca, and Polina Golland. Learning general-purpose biomedical volume representations using randomized synthesis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xOmC5LiVuN.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels. *arXiv preprint arXiv:2412.05496*, 2024.

Carmen Echávarri, P Aalten, Harry BM Uylings, HIL Jacobs, Pieter Jelle Visser, EHBM Gronenschild, FRJ Verhey, and S Burgmans. Atrophy in the parahippocampal gyrus as an early biomarker of alzheimer's disease. *Brain Structure and Function*, 215(3):265–271, 2011.

Brian L Edlow, Azma Mareyam, Andreas Horn, Jonathan R Polimeni, Thomas Witzel, M Dylan Tisdall, Jean C Augustinack, Jason P Stockmann, Bram R Diamond, Allison Stevens, et al. 7 tesla mri of the ex vivo human brain at 100 micron resolution. *Scientific data*, 6(1):244, 2019.

Andrea Esquivel, Andrea Ferrero, Achille Mileto, Francis Baffour, Kelly Horst, Prabhakar Shantha Rajiah, Akitoshi Inoue, Shuai Leng, Cynthia McCollough, and Joel G Fletcher. Photon-counting detector ct: key points radiologists should know. *Korean journal of radiology*, 23(9):854, 2022.

Fenja Falta, Christoph Großbröhmer, Alessa Hering, Alexander Bigalke, and Mattias P Heinrich. Lung250m-4b: A combined 3d dataset for CT- and point cloud-based intra-patient lung registration. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=FC0dsvguFi.

Miriam Friedel, Matthijs C van Eede, Jon Pipitone, M Mallar Chakravarty, and Jason P Lerch. Pydpiper: a flexible toolkit for constructing novel registration pipelines. *Frontiers in neuroinformatics*, 8:67, 2014.

Sarah F Frisken. Surfacenets for multi-label segmentations with preservation of sharp boundaries. *The Journal of computer graphics techniques*, 11(1):34, 2022.

Brett M Frye, Suzanne Craft, Thomas C Register, Jeongchul Kim, Christopher T Whitlow, Richard A Barcus, Samuel N Lockhart, Kiran Kumar Solingapuram Sai, and Carol A Shively. Early alzheimer's disease-like reductions in gray matter and cognitive function with aging in nonhuman primates. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 8(1):e12284, 2022.

Davide Gambarotto, Fabian U Zwettler, Maeva Le Guennec, Marketa Schmidt-Cernohorska, Denis Fortun, Susanne Borgers, Jörn Heine, Jan-Gero Schloetel, Matthias Reuss, Michael Unser, et al. Imaging cellular ultrastructures using expansion microscopy (u-exm). *Nature methods*, 16(1): 71–74, 2019.

James C Gee, Martin Reivich, and Ruzena Bajcsy. Elastically deforming a three-dimensional atlas to match anatomical brain images. 1993.

Hui Gong, Dongli Xu, Jing Yuan, Xiangning Li, Congdi Guo, Jie Peng, Yuxin Li, Lindsay A Schwarz, Anan Li, Bihe Hu, et al. High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level. *Nature communications*, 7(1):12142, 2016.

Maged Goubran, Cathie Crukley, Sandrine De Ribaupierre, Terence M Peters, and Ali R Khan. Image registration of ex-vivo mri to sparsely sectioned histology of hippocampal and neocortical temporal lobe specimens. *Neuroimage*, 83:770–781, 2013.

Courtney K Guo. *Multi-modal image registration with unsupervised deep learning*. PhD thesis, Massachusetts Institute of Technology, 2019.

Tripti Gupta, Gregory D Marquart, Eric J Horstick, Kathryn M Tabor, Sinisa Pajevic, and Harold A Burgess. Morphometric analysis and neuroanatomical mapping of the zebrafish brain. *Methods*, 150:49–62, 2018.

Alessa Hering, Keelin Murphy, and Bram van Ginneken. Learn2reg challenge: Ct lung registration - training data, 2020. URL https://doi.org/10.5281/zenodo.3835682.

Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712, 2022.

Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE transactions on medical imaging*, 41(3):543–558, 2021.

Junhao Hu, Weijie Gan, Zhixin Sun, Hongyu An, and Ulugbek S. Kamilov. A Plug-and-Play Image Registration Network, March 2024. URL http://arxiv.org/abs/2310.04297. arXiv:2310.04297 [eess].

Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H Adelson. Crisp boundary detection using pointwise mutual information. In *European conference on computer vision*, pp. 799–814. Springer, 2014.

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Reza Yazdani Aminadabi, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. System optimizations for enabling training of extreme long sequence transformer models. In *Proceedings of the 43rd ACM Symposium on Principles of Distributed Computing*, PODC '24, pp. 121–130, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706684. doi: 10.1145/3662158.3662806. URL https://doi.org/10.1145/3662158.3662806.

Rohit Jena, Pratik Chaudhari, and James C Gee. Fireants: Adaptive riemannian optimization for multi-scale diffeomorphic registration. *arXiv preprint arXiv:2404.01249*, 2024a.

Rohit Jena, Deeksha Sethi, Pratik Chaudhari, and James C. Gee. Deep learning in medical image registration: Magic or mirage? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 108331–108353. Curran Associates, Inc., 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c3fe2a07ec47b89c50e89706d2e23358-Paper-Conference.pdf.

Rohit Jena, Pratik Chaudhari, and James C. Gee. Deep implicit optimization enables robust learnable features for deformable image registration. *Medical Image Analysis*, 103:103577, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103577. URL https://www.sciencedirect.com/science/article/pii/S1361841525001240.

Xi Jia, Joseph Bartlett, Tianyang Zhang, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. U-net vs transformer: Is u-net outdated in medical image registration? *arXiv preprint arXiv:2208.04939*, 2022.

Xi Jia et al. A naive trick to accelerate training of lncc-based deep image registration models. *Preprints*, February 2025. doi: 10.20944/preprints202502.2200.v1.

Bailiang Jian, Jiazhen Pan, Morteza Ghahremani, Daniel Rueckert, Christian Wachinger, and Benedikt Wiestler. Mamba? catch the hype or rethink what really helps for image registration. *arXiv preprint arXiv:2407.19274*, 2024.

Hanna Jonsson, Simon Ekstrom, Robin Strand, Mette A Pedersen, Daniel Molin, Hakan Ahlstrom, and Joel Kullberg. An image registration method for voxel-wise analysis of whole-body oncological pet-ct. *Scientific Reports*, 12(1):18768, 2022.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Ankita Joshi and Yi Hong. Diffeomorphic Image Registration using Lipschitz Continuous Residual Networks. pp. 13.

Justin W Kenney, Patrick E Steadman, Olivia Young, Meng Ting Shi, Maris Polanco, Saba Dubaishi, Kristopher Covert, Thomas Mueller, and Paul W Frankland. A 3d adult zebrafish brain atlas (azba) for the digital age. *Elife*, 10:e69988, 2021.

Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, 29 (1):196–205, 2009.

David Kleinfeld, Arjun Bharioke, Pablo Blinder, Davi D Bock, Kevin L Briggman, Dmitri B Chklovskii, Winfried Denk, Moritz Helmstaedter, John P Kaufhold, Wei-Chung Allen Lee, et al. Large-scale automated histology in the pursuit of connectomes. *Journal of Neuroscience*, 31(45): 16125–16138, 2011.

Heidi Kleven, Ingvild E Bjerke, Francisco Clascá, Henk J Groenewegen, Jan G Bjaalie, and Trygve B Leergaard. Waxholm space atlas of the rat brain: a 3d atlas supporting data analysis and integration. *Nature methods*, 20(11):1822–1829, 2023.

Julian Krebs, Tommaso Mansi, Hervé Delingette, Li Zhang, Florin C Ghesu, Shun Miao, Andreas K Maier, Nicholas Ayache, Rui Liao, and Ali Kamen. Robust non-rigid registration through agent-based action learning. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pp. 344–352. Springer, 2017.

Fae N Kronman, Josephine K Liwang, Rebecca Betty, Daniel J Vanselow, Yuan-Ting Wu, Nicholas J Tustison, Ashwin Bhandiwad, Steffy B Manjila, Jennifer A Minteer, Donghui Shin, et al. Developmental mouse brain common coordinate framework. *Nature communications*, 15(1):9072, 2024.

Kwame S. Kutten, Joshua T. Vogelstein, Nicolas Charon, Li Ye, Karl Deisseroth M.D., and Michael I. Miller. Deformably registering and annotating whole CLARITY brains to an atlas via masked LDDMM. In Peter Schelkens, Touradj Ebrahimi, Gabriel Cristóbal, Frédéric Truchetet, and Pasi Saarikko (eds.), *Optics, Photonics and Digital Technologies for Imaging Applications IV*, volume 9896, pp. 989616. International Society for Optics and Photonics, SPIE, 2016. doi: 10.1117/12.2227444. URL https://doi.org/10.1117/12.2227444.

Joel Lamy-Poirier. Breadth-first pipeline parallelism. *Proceedings of Machine Learning and Systems*, 5:48–67, 2023.

Leo Lebrat, Rodrigo Santa Cruz, Frederic de Gournay, Darren Fu, Pierrick Bourgeat, Jurgen Fripp, Clinton Fookes, and Olivier Salvado. CorticalFlow: A Diffeomorphic Mesh Transformer Network for Cortical Surface Reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29491–29505. Curran Associates, Inc., 2021. URL https://papers.nips.cc/paper/2021/hash/f6b5f8c32c65fee991049a55dc97d1ce-Abstract.html.

Dacheng Li, Rulin Shao, Anze Xie, Eric P. Xing, Xuezhe Ma, Ion Stoica, Joseph E. Gonzalez, and Hao Zhang. DISTFLASHATTN: Distributed memory-efficient attention for long-context LLMs training. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=pUEDkZyPDl.

Shenggui Li, Fuzhao Xue, Chaitanya Baranwal, Yongbin Li, and Yang You. Sequence parallelism: Long sequence training from system perspective. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.134. URL https://aclanthology.org/2023.acl-long.134/.

Cher-Wei Liang, Ruey-Feng Chang, Pei-Wei Fang, and Chiao-Min Chen. Improving algorithm for the alignment of consecutive, whole-slide, immunohistochemical section images. *Journal of Pathology Informatics*, 12(1):29, 2021.

Jiayong Liang, Xiaoping Liu, Kangning Huang, Xia Li, Dagang Wang, and Xianwei Wang. Automatic registration of multisensor images using an integrated spatial and mutual information (smi) metric. *IEEE transactions on geoscience and remote sensing*, 52(1):603–615, 2013.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ringattention with blockwise transformers for near-infinite context. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=WsRHpHH4s0.

Yihao Liu, Junyu Chen, Lianrui Zuo, Aaron Carass, and Jerry L Prince. Vector field attention for deformable image registration. *Journal of Medical Imaging*, 11(6):064001–064001, 2024c.

Josephine K Liwang, Hannah C Bennett, Hyun-Jae Pi, and Yongsoo Kim. Protocol for using serial two-photon tomography to map cell types and cerebrovasculature at single-cell resolution in the whole adult mouse brain. *STAR protocols*, 4(1):102048, 2023.

Josephine K Liwang, Fae N Kronman, Hyun-Jae Pi, Yuan-Ting Wu, Daniel J Vanselow, Steffy B Manjila, Deniz Parmaksiz, Donghui Shin, Yoav Ben-Simon, Michael Taormina, et al. epdevatlas: mapping gabaergic cells and microglia in the early postnatal mouse brain. *Nature Communications*, 16(1):9538, 2025.

William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pp. 347–353. 1998.

Johannes Lotz, Janine Olesch, Benedikt Müller, Thomas Polzin, P Galuschka, JM Lotz, Stefan Heldmann, Hendrik Laue, Margarita González-Vallinas, Arne Warth, et al. Patch-based nonlinear image registration for gigapixel whole slide images. *IEEE Transactions on Biomedical Engineering*, 63(9):1812–1819, 2015.

Xin Luo, Zhigang Liu, Mingsheng Shang, Jungang Lou, and MengChu Zhou. Highly-accurate community detection via pointwise mutual information-incorporated symmetric non-negative matrix factorization. *IEEE Transactions on Network Science and Engineering*, 8(1):463–476, 2021. doi: 10.1109/TNSE.2020.3040407.

Falk Lüsebrink, Alessandro Sciarra, Hendrik Mattern, Renat Yakupov, and Oliver Speck. T1-weighted in vivo human whole brain mri dataset with an ultrahigh isotropic resolution of 250 μm. *Scientific data*, 4(1):1–12, 2017.

Mads AJ Madsen, Vanessa Wiggermann, Stephan Bramow, Jeppe Romme Christensen, Finn Sellebjerg, and Hartwig R Siebner. Imaging cortical multiple sclerosis lesions with ultra-high field mri. *NeuroImage: Clinical*, 32:102847, 2021.

Andreas Mang and Lars Ruthotto. A lagrangian gauss–newton–krylov solver for mass-and intensity-preserving diffeomorphic image registration. *SIAM Journal on Scientific Computing*, 39(5): B860–B885, 2017.

Andreas Mang, Amir Gholami, Christos Davatzikos, and George Biros. CLAIRE: A distributed-memory solver for constrained large deformation diffeomorphic image registration. *SIAM Journal on Scientific Computing*, 41(5):C548–C584, January 2019. ISSN 1064-8275, 1095-7197. doi: 10.1137/18M1207818. URL http://arxiv.org/abs/1808.04487. arXiv:1808.04487 [cs, math].

Harrison Mansour, Ryan Azrak, James J Cook, Kathryn J Hornburg, Yi Qi, Yuqi Tian, Robert W Williams, Fang-Cheng Yeh, Leonard E White, and G Allan Johnson. The duke mouse brain atlas: Mri and light sheet microscopy stereotaxic atlas of the mouse brain. *Science Advances*, 11(18): eadq8089, 2025.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. 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.

Gregory D Marquart, Kathryn M Tabor, Eric J Horstick, Mary Brown, Alexandra K Geoca, Nicholas F Polys, Damian Dalle Nogare, and Harold A Burgess. High-precision registration between zebrafish brain atlases using symmetric diffeomorphic normalization. *GigaScience*, 6(8):gix056, 2017.

David Mattes, David R Haynor, Hubert Vesselle, Thomas K Lewellyn, and William Eubank. Nonrigid multimodality image registration. In *Medical imaging 2001: image processing*, volume 4322, pp. 1609–1620. Spie, 2001.

Michael P Milham, Lei Ai, Bonhwang Koo, Ting Xu, Céline Amiez, Fabien Balezeau, Mark G Baxter, Erwin LA Blezer, Thomas Brochier, Aihua Chen, et al. An open resource for non-human primate imaging. *Neuron*, 100(1):61–74, 2018.

Tony C. W. Mok and Albert C. S. Chung. Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks, June 2020. URL http://arxiv.org/abs/2006.16148. arXiv:2006.16148 [cs, eess].

Tony CW Mok and Albert Chung. Affine medical image registration with coarse-to-fine vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20835–20844, 2022.

Keelin Murphy, Bram Van Ginneken, Joseph M Reinhardt, Sven Kabus, Kai Ding, Xiang Deng, Kunlin Cao, Kaifang Du, Gary E Christensen, Vincent Garcia, et al. Evaluation of registration methods on thoracic ct: the empire10 challenge. *IEEE transactions on medical imaging*, 30(11): 1901–1920, 2011.

Abdullah Nazib, James Galloway, Clinton Fookes, and Dimitri Perrin. Performance of registration tools on high-resolution 3d brain images. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 566–569, 2018. doi: 10.1109/EMBC.2018.8512403.

OpenAI. Triton. https://openai.com/index/triton/, 2021.

Hanchuan Peng, Phuong Chung, Fuhui Long, Lei Qu, Arnim Jenett, Andrew M Seeds, Eugene W Myers, and Julie H Simpson. Brainaligner: 3d registration atlases of drosophila brains. *Nature methods*, 8(6):493–498, 2011.

Zhen Peng, Minnan Luo, Wenbing Huang, Jundong Li, Qinghua Zheng, Fuchun Sun, and Junzhou Huang. Learning representations by graphical mutual information estimation and maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):722–737, 2023. doi: 10.1109/TPAMI.2022.3147886.

PyTorch. Fusing convolution and batch norm using custom function. https://docs.pytorch.org/tutorials/intermediate/custom_function_conv_bn_tutorial.html, 2023. Created July 22, 2021; Last updated April 18, 2023; Last verified November 5, 2024.

PyTorch. Helion. https://pytorch.org/blog/helion/, 2025.

Penghui Qi, Xinyi Wan, Guangxing Huang, and Min Lin. Zero bubble pipeline parallelism. *arXiv preprint arXiv:2401.10241*, 2023.

Huaqi Qiu, Chen Qin, Andreas Schuh, Kerstin Hammernik, and Daniel Rueckert. Learning diffeomorphic and modality-invariant registration using b-splines. 2021.

Quan Quan, Qingsong Yao, Heqin Zhu, and S Kevin Zhou. Igu-aug: Information-guided unsupervised augmentation and pixel-wise contrastive learning for medical image analysis. *IEEE Transactions on Medical Imaging*, 2024.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16, 2020. doi: 10.1109/SC41405.2020.00024.

Sadhana Ravikumar, Laura E. M. Wisse, Sydney Lim, Ranjit Ittyerah, Long Xie, Madigan L. Bedard, Sandhitsu R. Das, Edward B. Lee, M. Dylan Tisdall, Karthik Prabhakaran, Jacqueline Lane, John A. Detre, Gabor Mizsei, John Q. Trojanowski, John L. Robinson, Theresa Schuck, Murray Grossman, Emilio Artacho-Pérula, Maria Mercedes Iñiguez de Onzoño Martin, María del Mar Arroyo Jiménez, Monica Muñoz, Francisco Javier Molina Romero, Maria del Pilar Marcos Rabal, Sandra Cebada Sánchez, José Carlos Delgado González, Carlos de la Rosa Prieto, Marta Córcoles Parada, David J. Irwin, David A. Wolk, Ricardo Insausti, and Paul A. Yushkevich. Ex vivo mri atlas of the human medial temporal lobe: characterizing neurodegeneration due to tau pathology. *Acta Neuropathologica Communications*, 9(1):173, 2021. ISSN 2051-5960. doi: 10.1186/s40478-021-01275-7. URL https://doi.org/10.1186/s40478-021-01275-7.

Sadhana Ravikumar, Amanda E Denning, Sydney Lim, Eunice Chung, Niyousha Sadeghpour, Ranjit Ittyerah, Laura EM Wisse, Sandhitsu R Das, Long Xie, John L Robinson, et al. Postmortem imaging reveals patterns of medial temporal lobe vulnerability to tau pathology in alzheimer's disease. *Nature Communications*, 15(1):4803, 2024.

Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: learning deformable image registration using shape matching. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pp. 266–274. Springer, 2017.

Suman Sarkar and Biswajit Pandey. A study on the statistical significance of mutual information between morphology of a galaxy and its large-scale environment. *Monthly Notices of the Royal Astronomical Society*, 497(4):4077–4090, 2020.

Will Schroeder and Spiros Tsalikis. Really fast isocontouring. https://www.kitware.com/really-fast-isocontouring/, June 13 2023. Kitware blog / announcement.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Henrik Skibbe, Muhammad Febrian Rachmadi, Ken Nakae, Carlos Enrique Gutierrez, Junichi Hata, Hiromichi Tsukada, Charissa Poon, Matthias Schlachter, Kenji Doya, Piotr Majka, et al. The brain/minds marmoset connectivity resource: An open-access platform for cellular-level tracing and tractography in the primate brain. *PLoS biology*, 21(6):e3002158, 2023.

Hessam Sokooti, Bob De Vos, Floris Berendsen, Boudewijn PF Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pp. 232–239. Springer, 2017.

Benjamin Frederick Spector, Simran Arora, Aaryan Singhal, Arjun Parthasarathy, Daniel Y Fu, and Christopher Re. Thunderkittens: Simple, fast, and $\textit{Adorable}$ kernels. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0fJfVOSUra.

Bane Sullivan and Alexander Kaszynski. PyVista: 3D plotting and mesh analysis through a streamlined interface for the Visualization Toolkit (VTK). *Journal of Open Source Software*, 4(37):1450, May 2019. doi: 10.21105/joss.01450. URL https://doi.org/10.21105/joss.01450.

Kathryn M Tabor, Gregory D Marquart, Christopher Hurt, Trevor S Smith, Alexandra K Geoca, Ashwin A Bhandiwad, Abhignya Subedi, Jennifer L Sinclair, Hannah M Rose, Nicholas F Polys, et al. Brain-wide cellular resolution imaging of cre transgenic zebrafish lines for functional circuit-mapping. *Elife*, 8:e42687, 2019.

Nouamane Tazi, Ferdinand Mom, Haojun Zhao, Phuc Nguyen, Mohamed Mekkouri, Leandro Werra, and Thomas Wolf. The ultra-scale playbook: Training LLMs on GPU clusters, 2024. URL https://huggingface.co/blog/the-ultra-scale-playbook. HuggingFace Blog.

Philippe Thévenaz and Michael Unser. Optimization of mutual information for multiresolution image registration. *IEEE transactions on image processing*, 9(12):2083–2099, 2000.

Lin Tian, Hastings Greer, Roland Kwitt, François-Xavier Vialard, Raúl San José Estépar, Sylvain Bouix, Richard Rushmore, and Marc Niethammer. unigradicon: A foundation model for medical image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 749–760. Springer, 2024.

Lazaros C Triarhou. Dopamine and parkinson's disease. In *Madame curie bioscience database [internet]*. Landes Bioscience, 2013.

Nicholas J Tustison and Brian B Avants. Explicit b-spline regularization in diffeomorphic image registration. *Frontiers in neuroinformatics*, 7:39, 2013.

Nicholas J Tustison, Min Chen, Fae N Kronman, Jeffrey T Duda, Clare Gamlin, Mia G Tustison, Michael Kunst, Rachel Dalley, Staci Sorenson, Quanxin Wang, et al. The antsx ecosystem for mapping the mouse brain. *bioRxiv*, pp. 2024–05, 2024.

Erdem Varol, Amin Nejatbakhsh, Ruoxi Sun, Gonzalo Mena, Eviatar Yemini, Oliver Hobert, and Liam Paninski. Statistical atlas of c. elegans neurons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pp. 119–129. Springer, 2020.

Vivek Venkatachalam, Ni Ji, Xian Wang, Christopher Clark, James Kameron Mitchell, Mason Klein, Christopher J Tabone, Jeremy Florman, Hongfei Ji, Joel Greenwood, et al. Pan-neuronal imaging in roaming caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 113(8): E1082–E1088, 2016.

Guoxia Wang, Jinle Zeng, Xiyuan Xiao, Siming Wu, Jiabin Yang, Lujing Zheng, Zeyu Chen, Jiang Bian, Dianhai Yu, and Haifeng Wang. Flashmask: Efficient and rich mask extension of flashattention. *arXiv preprint arXiv:2410.01359*, 2024.

Quanxin Wang, Song-Lin Ding, Yang Li, Josh Royall, David Feng, Phil Lesnar, Nile Graddis, Maitham Naeemi, Benjamin Facer, Anh Ho, Tim Dolbeare, Brandon Blanchard, Nick Dee, Wayne Wakeman, Karla E. Hirokawa, Aaron Szafer, Susan M. Sunkin, Seung Wook Oh, Amy Bernard, John W. Phillips, Michael Hawrylycz, Christof Koch, Hongkui Zeng, Julie A. Harris, and Lydia Ng. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell*, 181(4):936–953.e20, May 2020a. ISSN 00928674. doi: 10.1016/j.cell.2020.04.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867420304025.

Quanxin Wang, Song-Lin Ding, Yang Li, Josh Royall, David Feng, Phil Lesnar, Nile Graddis, Maitham Naeemi, Benjamin Facer, Anh Ho, et al. The allen mouse brain common coordinate framework: a 3d reference atlas. *Cell*, 181(4):936–953, 2020b.

Asmamaw T Wassie, Yongxin Zhao, and Edward S Boyden. Expansion microscopy: principles and uses in biological research. *Nature methods*, 16(1):33–41, 2019.

Thomas Welton, Septian Hartono, Yao-Chia Shih, Stefan T Schwarz, Yue Xing, Eng-King Tan, Dorothee P Auer, Noam Harel, and Ling-Ling Chan. Ultra-high-field 7t mri in parkinson's disease: ready for clinical use?—a narrative review. *Quantitative Imaging in Medicine and Surgery*, 13(11): 7607, 2023.

Marek Wodzinski, Niccolo Marini, Manfredo Atzori, and Henning Müller. Deeperhistreg: robust whole slide images registration framework. *arXiv preprint arXiv:2404.14434*, 2024.

Yifan Wu, Tom Z. Jiahao, Jiancong Wang, Paul A. Yushkevich, M. Ani Hsieh, and James C. Gee. NODEO: A Neural Ordinary Differential Equation Based Optimization Framework for Deformable Image Registration. *arXiv:2108.03443 [cs]*, February 2022. URL http://arxiv.org/abs/2108.03443. arXiv: 2108.03443.

Yifan Wu, Mengjin Dong, Rohit Jena, Chen Qin, and James C Gee. Neural ordinary differential equation based sequential image registration for dynamic characterization. *arXiv preprint arXiv:2404.02106*, 2024.

Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158:378–396, 2017.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, YX Wei, Lean Wang, Zhiping Xiao, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. *arXiv preprint arXiv:2502.11089*, 2025.

Liutong Zhang, Lei Zhou, Ruiyang Li, Xianyu Wang, Boxuan Han, and Hongen Liao. Cascaded feature warping network for unsupervised medical image registration. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 913–916. IEEE, 2021.

Shengyu Zhao, Yue Dong, Eric I-Chao Chang, and Yan Xu. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019a.

Shengyu Zhao, Tingfung Lau, Ji Luo, I Eric, Chao Chang, and Yan Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network. *IEEE journal of biomedical and health informatics*, 24(5):1394–1404, 2019b.

Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019c.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.