Orochi: Versatile Biomedical Image Processor

Gaole Dai[1,*] Chenghao Zhou[1,2,*] Yu Zhou[3,*]

Rongyu Zhang[1] Yuan Zhang[1] Chengkai Hou[1] Tiejun Huang[1]

Jianxu Chen[3,✉]
jianxu.chen@isas.de

Shanghang Zhang[1,✉]
shanghang@pku.edu.cn

[1] State Key Laboratory of Multimedia Information Processing,
School of Computer Science, Peking University
[2] Academy for Advanced Interdisciplinary Studies, Peking University
[3] Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V.

Abstract

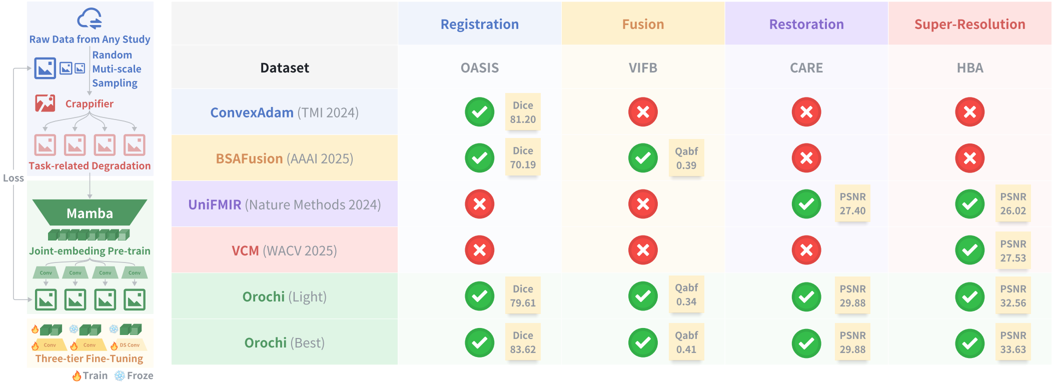
Deep learning has emerged as a pivotal tool for accelerating research in the life sciences, with the low-level processing of biomedical images (e.g., registration, fusion, restoration, super-resolution) being one of its most critical applications. Platforms such as ImageJ (Fiji) and napari have enabled the development of customized plugins for various models. However, these plugins are typically based on models that are limited to specific tasks and datasets, making them less practical for biologists. To address this challenge, we introduce **Orochi**, the first application-oriented, efficient, and versatile image processor designed to overcome these limitations. Orochi is pre-trained on patches/volumes extracted from the raw data of over 100 publicly available studies using our Random Multi-scale Sampling strategy. We further propose Task-related Joint-embedding Pre-Training (TJP), which employs biomedical task-related degradation for self-supervision rather than relying on Masked Image Modelling (MIM), which performs poorly in downstream tasks such as registration. To ensure computational efficiency, we leverage Mamba's linear computational complexity and construct Multi-head Hierarchy Mamba. Additionally, we provide a three-tier fine-tuning framework (Full, Normal, and Light) and demonstrate that Orochi achieves comparable or superior performance to current state-of-the-art specialist models, even with lightweight parameter-efficient options. We hope that our study contributes to the development of an all-in-one workflow, thereby relieving biologists from the overwhelming task of selecting among numerous models. Our pre-trained weights and code will be released.

1 Introduction

With the rapid advancement of deep learning, modern neural networks have demonstrated remarkable scalability and spawned a wide array of downstream applications in AI for Life Science [1, 2, 3]. Among these, biomedical image processing is a pivotal topic. Its significance arises from the inherent

[*] Equal Contribution, ✉Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



	Registration	Fusion	Restoration	Super-Resolution
Dataset	OASIS	VIFB	CARE	HBA
ConvexAdam (TMI 2024)	✓ Dice 81.20	✗	✗	✗
BSAFusion (AAAI 2025)	✓ Dice 70.19	✓ Qabf 0.39	✗	✗
UniFMIR (Nature Methods 2024)	✗	✗	✓ PSNR 27.40	✓ PSNR 26.02
VCM (WACV 2025)	✗	✗	✗	✓ PSNR 27.53
Orochi (Light)	✓ Dice 79.61	✓ Qabf 0.34	✓ PSNR 29.88	✓ PSNR 32.56
Orochi (Best)	✓ Dice 83.62	✓ Qabf 0.41	✓ PSNR 29.88	✓ PSNR 33.63

Figure 1: **Trend of Versatile Biomedical Image Precessor.** We listed the recent advancements in biomedical image processing, where matched row-to-column colour coding highlights the main task of each model. Stickers display the reported scores from the respective papers. Orochi extends the versatile bandwidth and exhibits exceptional performance across tasks and tuning modes.

constraints in acquiring biomedical images compared to natural images, which often compromise source image quality. Specifically, the most common limitations stem from imaging device operational trade-offs. For instance, in optical microscopy, excessive laser intensity can damage target tissues, while insufficient laser power introduces low signal-to-noise ratios [4]. Similarly, in computed tomography (CT), thinner slice scans subject patients to prolonged high radiation exposure, posing health risks, whereas sparse slicing results in low-resolution data [5]. These challenges drive the demand for biomedical image **restoration** [6, 4, 7, 8, 9] and **super-resolution** [10, 11, 12, 13, 5] tasks. Another class of limitations originates from the intrinsic shortcomings of imaging modalities. For example, CT imaging is efficient and provides clear hierarchical information but suffers from poor soft-tissue contrast, in contrast, magnetic resonance imaging (MRI) excels in soft-tissue resolution but requires longer acquisition times and is susceptible to motion artifacts. Such modality-specific weaknesses necessitate biomedical image **fusion** tasks [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. Furthermore, acquiring synchronous multi-modal data imposes demands on equipment and environments, making asynchronous data more prevalent. However, misalignment exists between tissues or cells due to temporal/specimen variability, motivating image **registration** [26, 27, 28, 29, 30, 31, 32, 33] tasks to align asynchronous or even heterogeneous datasets of the same specimen.

With the emergence of long-range dependency models [34, 35, 36] and self-supervised pre-training methods [37, 38, 39, 40], models designed for the aforementioned issues have advanced rapidly, giving rise to powerful specialist models (see Figure 1). However, we argue that in practical applications, these specialist models neglect three critical factors: **(1) Task Perspective:** Real-world biomedical imaging tasks often require multiple sequential steps (e.g., registration followed by fusion, as discussed earlier). **(2) Degradation Perspective:** Since the underlying causes of degradation share similarities, these degradations are interrelated—for example, both low signal-to-noise ratio and low resolution result in information loss. **(3) Data Perspective:** Due to their characteristics of being multi-channel, large-scale, and high-throughput, biomedical images are considerably larger than natural images, making the training and inference of multiple specialist models highly inefficient. From both efficiency and effectiveness standpoints, these issues collectively motivate the development of a universal foundational model. We aim for such a generalist model to optimize the aforementioned challenges by: **(1)** handling diverse low-level tasks within a unified framework, thereby avoiding the difficulties of selecting and integrating several specialist models; **(2)** capturing more generalized and robust features via cross-task learning during the pre-training phase; and **(3)** addresses real-world biomedical data processing costs to reduce redundant training and inference.

Therefore, we introduce **Orochi** (named after the legendary multi-headed serpent). To fulfill the envisioned goals, our design emphasizes four aspects (see Figure 2): **(1) Dataset Level:** We extensively employ unlabeled raw data from over 100 publicly available studies (see Appendix A.2) and perform our Random Multi-scale Sampling, which considers the different scales of Region-of-Interest (ROI). **(2) Pre-training Level:** Inspired by Joint-embedding Prediction Architecture (JEPA) [40], where different degradations serve as context for each others. Our Task-related Joint-

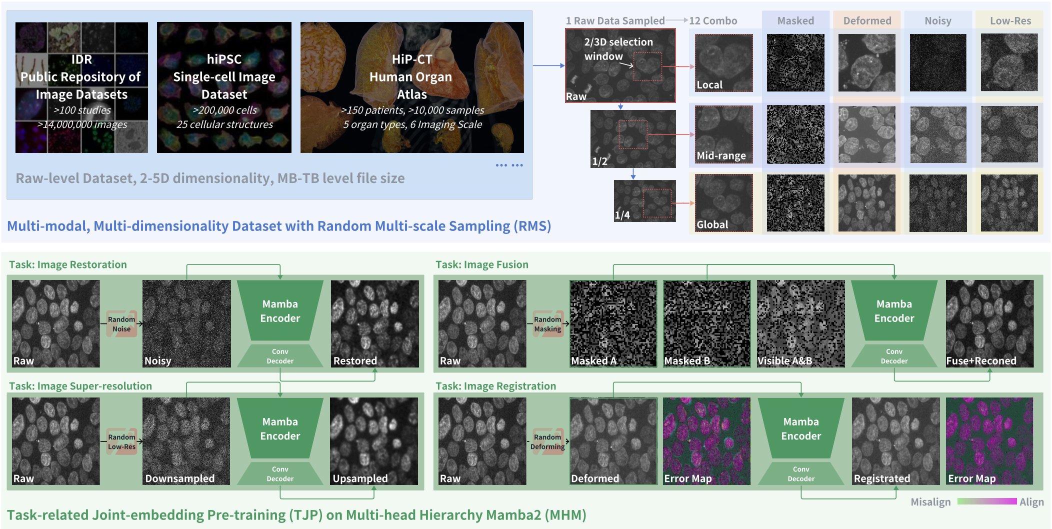


Figure 2: **Overview of the Construction of Orochi.** The upper panel illustrates the data conversion pipeline, taking into account patches/volumes at multiple scales. The lower panel presents the self-supervised strategies utilized during pre-training. Additionally, we provide supplementary images (e.g., Visible A+B, Error Map) to facilitate the comparison between inputs and outputs.

embedding Pre-training (TJP) applies various forms of task-specific degradation, and the model learns from reconstructing them jointly. **(3) Model Level:** On one hand, we employ Mamba [41] as the building blocks to leverages its linear complexity [42, 41, 43, 44]. On the other hand, the overall structure draws inspiration from the hierarchical design of the Swin-Transformer [36] by incorporating patch merging to enhance model efficiency further. **(4) Post-training Level:** We propose a three-tier fine-tuning framework to reduce the tuning cost. Ranging from full fine-tuning (Full), to fine-tuning only the replaced dense convolution head (Normal), and finally to the most lightweight variant using depth-wise separable convolution [45] (Light), thereby achieving Parameter-Efficient Fine-Tuning (PEFT) [46]. To this end, we hope that Orochi will distinguish itself as an exceptional tool among the extensive array of plugins available on platforms such as ImageJ (Fiji) [47] and napari [48], further advancing towards a user-friendly workflow with unified functionalities.

In summary, our main contributions are as follows:

1. We systematically review the significance of low-level biomedical image processing and highlight that even in this era of powerful foundational models, the paradigm centred on specialist models still exhibits inherent deficiencies. Limiting both effectiveness and efficiency from the perspectives of task, degradation, and data. **To the best of our knowledge, Orochi is the first versatile foundational model addressing these issues.**

2. We curated raw-level datasets from over 100 studies [49, 50, 51], covering a wide range of imaging modalities from 2-5D — with a total data size over 100 terabytes. During training, we introduce Random Multi-scale Sampling to achieve a unitedly raw data conversion into training patches/volumes. These converted data are used for both local and stream training, alleviating the challenges with the transmission and storage of extremely large datasets.

3. We propose Task-related Joint-embedding Pre-training (TJP), which directly learns the inter-relations among various task-specific degradations rather than relying on common Masked Image Modelling (MIM). For the model architecture, we leverage the linear complexity of Mamba and design a multi-head hierarchical structure to minimize the costs of training and inference. Finally, for post-training, we introduce a three-tier fine-tuning framework and demonstrate that even the most lightweight depth-separable convolution tuning can achieve performance comparable to existing state-of-the-art specialist models.

2 Related Works

Self-supervised Learning Self-supervised Learning (SSL) extracts inherent data properties. Masked Image Modelling (MIM) predicts masked image regions from original pixel values using an encoder-decoder architecture, with loss in image space [37, 38]. Contrastive Learning (CL) aligns representations of augmented views of the same image in an embedding space via specialized objectives [52, 39]. Combining these, the Joint-Embedding Predictive Architecture (JEPA) [40] predicts full latent representations from context to learn robust image representations.

Restoration To address low image quality in fluorescence microscopy, Content-aware image restoration (CARE) [4] uses CNNs. Li *et al.* [8] improved axial resolution using a CARE-based model with physically acquired ground truth. Subsequent works integrated Swin-Transformers [36] for efficiency (SwinIR [9]) or Mamba blocks [42] for long-range dependency modeling (MambaIR [7]). UniFMIR [6] demonstrated that pre-trained foundation models generalize well for this task.

Super-resolution Super-resolution aims to overcome optical limits. DeepLP [53] employs point-scanning for reconstruction. Diffusion-based models, including volumetric conditioning modules [5] and latent diffusion in InverseSR [11], show promise for 3D brain MRI. Other approaches include local implicit image functions for flexible resolution enhancement [13], joint super-resolution and synthesis frameworks for isotropic volumes [12], and methods for multimodal image super-resolution [10].

Registration Image registration aligns images by optimizing a deformation field. VoxelMorph [54] provides a learning-based 3D framework. Dual-encoder U-Nets [26], Swin-Transformers for long-distance correspondences (TransMorph [32]), and Mamba blocks [42] for efficient long-range modeling (MambaMorph [33]). Fast 3D registration methods have been proposed by Siebert *et al.* [27, 29], while Mok *et al.* [30, 28] address large deformations with Laplacian Pyramid Networks.

Fusion Multi-modality image fusion integrates complementary information. Techniques include bidirectional stepwise feature alignment for unaligned images (BSAFusion [25]), mutual enhancement for PAT/MRI fusion [24], and diffusion-based methods incorporating fusion priors (Diff-IF [55]) or denoising diffusion models [16]. Semantic-aware strategies with registration are found in SuperFusion [21] and MURF [22]. Other notable methods encompass one-stage progressive dense registration [23], U2Fusion [14], and Equivariant fusion [15]. Diverse strategies also include lightweight and semantic-guided approaches (ALMFNET [17], MSGFUSION [18]), dictionary-based and GAN-driven frameworks [19, 56], and unsupervised methods [20].

3 Methods

Due to page limitations, this section primarily emphasizes our comprehensive degradation designs used for self-supervision. The Appendix provided detailed architecture of the Multi-head Hierarchy Mamba model along with the three-tier fine-tuning framework B.

3.1 Preliminary: Self-supervised Degradation

Self-supervised image learning can be generally formulated as learning a reconstruction function f_θ that recovers the original image x from its degraded $D(x)$. Formally, this objective is defined as:

$$\min_{\theta} \mathbb{E}_x \left[\ell\Big(x, f_\theta\big(D(x)\big)\Big) \right], \tag{1}$$

where x is the sampled data, p_{data}, $D(\cdot)$ denotes a degradation function applied to x, f_θ is the parameterized model, and ℓ is a loss function (e.g., the L2 loss or perceptual loss).

Masked Image For masked image degradation, the degradation function is defined as: $D_{\text{mask}}(x) = x \odot M$, where $M \in \{0,1\}^{H \times W}$ is a binary mask with height H and width W that selectively occludes regions of x. This degradation helps the model learn to infer missing information.

Deformed Image For deformed image degradation, the degradation function takes the form: $D_{\text{def}}(x) = \mathbf{T}(x)$, where $\mathbf{T}(\cdot)$ represents a spatial transformation (such as rotation, scaling, or warping). This degradation introduces geometric distortions that mimic real-world variations.

Nosiy Image For noisy image degradation, the degradation function is defined as: $D_{\text{noise}}(x) = x + \eta$ where η denotes additive noise (typically Gaussian noise), simulating sensor imperfections or environmental interference.

Low-resolution Image For low-resolution image degradation, the degradation function is given by: $D_{\text{LR}}(x) =\downarrow_s (x)$, where $\downarrow_s$ is a down-sampling operator with scale factor s, reducing the resolution of x to simulate the effects of low-resolution imaging.

3.2 Orochi: Random Multi-scale Sampling

Random Multi-scale Sampling aims to extract patches/volumes with diverse scales from raw images. Given a raw image I, the procedure consists of two main steps: **(1) Multi-scale Resizing:** We first generate scaled versions of the raw image I to capture features at different resolutions. In particular, we resize I to scales $1/2$ and $1/4$ of its original size. Formally, let:

$$I_s =\downarrow_s (I), \quad s \in \{1, \tfrac{1}{2}, \tfrac{1}{4}\}, \tag{2}$$

where $\downarrow_s (\cdot)$ denotes down-sampling with factor s. **(2) Random Window Sampling:** For each scaled image I_s, we define a fixed-size window K (compatible with the pre-training requirements in either 2D or 3D) and perform random sampling to extract sub-patches. Let the window K have dimensions $W \times H$ (or $W \times H \times D$ for 3D data). A randomly sampled 2D patch x_s at scale s is given by:

$$x_s = I_s\big(i : i + W - 1,\ j : j + H - 1\big), \tag{3}$$

where $(i,\ j)$ is a randomly chosen starting coordinate in I_s.

Collectively, the set of patches extracted across scales is represented as:

$$x = \{x_{s,n} \mid s \in \{1, \tfrac{1}{2}, \tfrac{1}{4}\},\ n = 1, \ldots, \mathrm{N}_s\}, \tag{4}$$

where N_s denotes the number of patches sampled from the image at scale s. These multi-scale patches are then passed to subsequent degradation processes (e.g., masking, deformation, noise addition, and low-resolution conversion). By performing random sampling across multiple scales, our method extended the data diversity and enabled more robust feature learning across various datasets.

3.3 Orochi: Task-related Joint-embedding Pre-training

Dual-Masking Reconstructive Fusion To better address the biomedical image fusion task, where the combination of existing contexts is crucial, we modified the conventional Masked Image Modelling approaches [37, 38], which typically employ a single masking strategy. Specifically, we applied two distinct masking operations to the training data x, thereby generating two independent masks:

$$x_A = x \odot M_A, \quad x_B = x \odot M_B, \tag{5}$$

where $M_A, M_B \in \{0,1\}^{H \times W}$ are binary masks with only partial overlap and ensure invisible information retention even after fusion. The masking probabilities are generated by:

$$M_k[i,j] = \mathbf{1}[\xi^k_{i,j} < \tau], \quad k \in A, B, \tag{6}$$

where $\xi^k_{i,j} \sim \mathbf{U}(0,1)$ represents a random value extracted from a uniform distribution for grid coordinates i, j, and τ is the masking threshold. The key innovation is that our model is exposed to process both masked inputs (x_A, x_B) simultaneously to recover the original image: $\hat{x} = f_\theta(x_A, x_B)$,. This guides the model to develop robust feature extraction capabilities that can identify complementary information across different masked views, and then fuse these partial observations coherently to reconstruct missing regions in both inputs.

Spatially-varying Gaussian down-sample For down-sampling, we adapt similar principles from DeepLP [53], which tested noisy down-sampling beyond uniform down-sampling in self-supervised microscopy restoration. We enhance this noisy down-sampling with spatially varying characteristics:

$$D_{\text{LR}}(x) = \mathbf{G}\sigma\text{var}(\uparrow_{\frac{1}{s}} (\downarrow_s (x + \eta))), \tag{7}$$

where $\downarrow_s$ represents down-sampling with a random scale factor s, $\uparrow_{\frac{1}{s}}$ denotes upsampling back to the original resolution, $\eta \sim \mathbf{N}(0, \sigma^2_{\text{down}})$ is normal distributed noise added during the down-sampling process with $\sigma_{\text{down}} \sim \mathbf{U}(0.01, 0.1)$, $\mathbf{U}$ represent uniform distribution, and $\mathbf{G}\sigma$var denotes spatially-varying Gaussian filtering. It can be defined as:

$$\mathbf{G}\sigma\text{var}(x)[i,j] = \sum_{u,v} g_{\sigma(i,j)}(u,v) \cdot x[i-u, j-v], \tag{8}$$

where g_σ represents a Gaussian kernel (2/3D) with standard deviation $\sigma(i,j) \sim \mathbf{U}(\sigma_{\text{min}}, \sigma_{\text{max}})$ that varies across grid coordinates i, j. This mimics the heterogeneous blurring found in optical systems.

Multi-scale Smoothed Perlin Noise Deformation For the self-supervised registration task, constructing a realistic deformation field is important. We conducted multi-scale Perlin noise fields that simulate the hierarchy variations in natural anatomical structures. Given an image x, we generate a deformation field Φ and its corresponding deformed image $D_{\text{def}}(x)$ as follows:

$$D_{\text{def}}(x) = \mathbf{T}(x, \Phi), \quad \Phi = \mathbf{G}_\sigma(\mathbf{Per}(\mathbf{f}, \mathbf{p})), \tag{9}$$

$\mathbf{T}(\cdot,\cdot)$ is a spatial transformation operator, $\mathbf{G}_\sigma(\cdot)$ denotes spatially-varying Gaussian smoothing with parameter σ, and $\mathbf{Per}(\mathbf{f}, \mathbf{p})$ represents multi-octave Perlin noise with frequency $\mathbf{f}$ and persistence $\mathbf{p}$.

The multi-octave Perlin noise is specifically defined as:

$$\mathbf{Per}(\mathbf{f}, \mathbf{p}) = \sum_{n=1}^{\text{N}} \mathbf{p}^{n-1} \cdot \mathbf{S}(\mathbf{f}^{n-1} \cdot (i,j)), \tag{10}$$

where $\mathbf{S}(\cdot)$ is the simplex noise function, N is the number of octaves and **coords** represents the grid coordinates. This multi-scale approach generates deformation fields with varying levels of detail.

To enhance the anatomical plausibility of the deformations, we apply normalization and bound it using a tanh function: $\Phi_{\text{final}} = \alpha \cdot \tanh(\Phi)$, where α controls the maximum displacement magnitude.

Multi-stage Noise Simulation To simulate realistic noise, we adopted a multi-stage process:

$$D_{\text{noise}}(x) = \mathbf{Bi}_p(\mathbf{Poi}(\max(0, x+\eta))), \tag{11}$$

where $\eta \sim \mathbf{N}(0, \sigma^2_{\text{noise}})$ with $\sigma_{\text{noise}} \sim \mathbf{U}(0.075, 0.15)$ represents Gaussian noise, $\mathbf{Poi}(\lambda)$ denotes Poisson noise with intensity parameter λ (modeling photon-counting statistics), and $\mathbf{Bi}_p$ represents binary (salt-and-pepper) noise that affects a proportion of pixels with probability p.

These sophisticated degradation designs enable our framework to simulate a wide spectrum of real-world imaging artifacts, encouraging the model to handle diverse image quality issues encountered.

4 Experiments

We conducted comprehensive comparisons strictly following the setups in published specialist models (UniFMIR [6], VCM [5], Transmorph [32], and BSAFusion [25], see Appendix A.4 for details). Resulting in more than 30 state-of-the-art baselines across multiple benchmarks for various biomedical image-processing tasks to demonstrate the effectiveness and versatility of Orochi. We color-coded the performance in Table 1, 2, 3, and 4 with **Red (1st)**, Blue (2nd), and the row color reflects the training type with [Training-free], [Training from Scratch], [Fine-Tuning], and [Efficent Fine-Tuning]. See the Appendix for more details of the experiment setups A and extra validation C.

Generalization Capability on In-Domain Data Given that our model, Orochi, is extensively pretrained, we expect it to exhibit strong generalization capabilities on in-domain data. Accordingly, in Figure 3 we demonstrate Orochi's zero-shot performance on various stained microscopy images [51] (results on clinical images [50] are detailed in the Appendix C.1). Panels (A)–(D) illustrate Orochi's robust processing capabilities. In Panel (E), we further examine whether these outcomes align with our algorithmic expectations. For example, our Dual-Masking Reconstructive Fusion anticipates that the model learns an effective fusion strategy and leverages the existing information from both

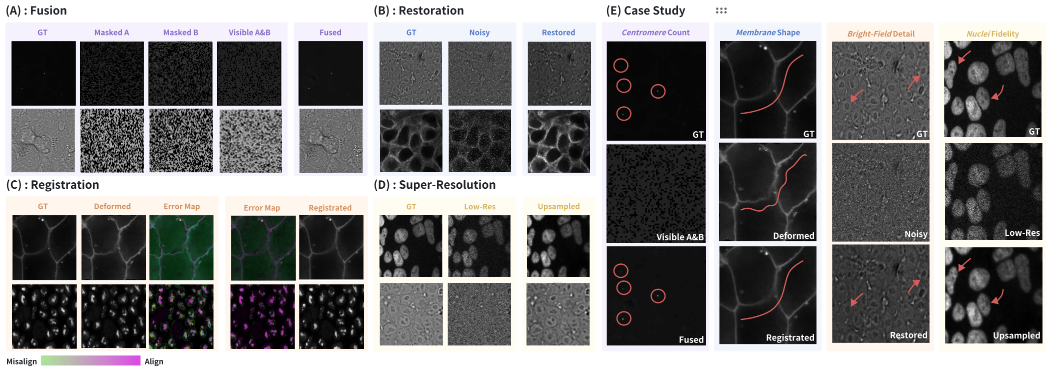


Figure 3: **In-Domain Generalization Performance of Orochi on Unseen Test Images.** (A)–(D) illustrate Orochi's robust performance across various low-level processing tasks when applied to unseen testing images after pre-training. Supplementary images include the dual-masking images and naive merge results for the fusion task. Error maps for the registration task. (E) provides in-depth case studies: for the fusion task, the centromere count is emphasized in both the reconstructed image and the original image (highlighted with circles); for the registration task, subtle deformations of the cell membrane are accentuated; and for the restoration and super-resolution tasks, the fine details of bright-field images and the internal structures of DNA-stained cell nuclei are emphasized

Table 1: **Isotropic 3D volume Restoration Task.** Low-high laser data pairs along the XY axis are collected and serve as the training set. However, the evaluation is on both XY slices and XZ slices.

Method	**PSNR (XY)** ↑	**SSIM (XY)** ↑	**PSNR (XZ)** ↑	**SSIM (XZ)** ↑
Dataset: ***CARE*** **[4]**				
Li *et al.* [8]	23.71	0.58	24.51	0.58
CARE [4]	25.60	0.60	25.76	0.64
SwinIR [9]	25.98	0.62	26.41	0.64
MambaIR [7]	25.89	0.64	27.17	0.65
UniFMIR [6]	27.12	0.66	27.67	0.66
prune-UniFMIR (FP16) [6]	25.00	0.59	26.48	0.64
prune-UniFMIR (FP32) [6]	26.18	0.63	26.24	0.64
Orochi (Full)	28.31	0.70	28.52	0.71
Orochi (Normal)	29.15	**0.71**	29.43	0.71
Orochi (Light)	**29.77**	**0.71**	**29.98**	**0.72**

Table 2: **MRI Axial Super-resolution Task.** Two intensities of low-resolution data are trained and tested in this task, with 4mm (i.e x4 down-sampled) and 8mm (i.e x8 down-sampled)

Method	**PSNR (4mm)** ↑	**SSIM (4mm)** ↑	**PSNR (8mm)** ↑	**SSIM (8mm)** ↑
Dataset: ***HBA*** **[57]**				
Cubic	23.84	0.76	21.80	0.63
UniRes [10]	21.49	0.69	20.91	0.63
SynthSR [12]	19.22	0.66	19.02	0.62
LIIF [13]	32.41	0.95	25.12	0.81
InverseSR-LDM [11]	28.59	0.80	27.92	0.75
InverseSR [11]	27.51	0.88	23.66	0.79
VCM [5]	27.54	0.87	27.52	0.86
Orochi (Full)	**35.33**	0.95	**31.93**	0.89
Orochi (Normal)	34.60	**0.96**	29.51	0.89
Orochi (Light)	34.83	**0.96**	30.28	**0.90**

sources before reconstruction, rather than reconstructing masked regions separately. This expectation is validated in the *Centromere* count case study, where masked A and B each exhibit a partial absence of centromeres, and the **model successfully performs complementary fusion and the final reconstruction does not arbitrarily generate *Centromeres* across extensive background regions.** This precise control over fine structural details is also evident in other cases.

Image Restoration Task In Table 1, we present the performance of Orochi on the isotropic 3D volume restoration task. In microscopy imaging, the image quality along the XY plane is typically much higher than that along the XZ plane due to the inherent limitations of sequential (layer-by-layer) imaging, such as in light-sheet microscopy, which leads to the formation of isotropic data. To address this, CARE [4] leverages the high-resolution XY data for training and subsequently restores the lower-resolution XZ data. On this task, Orochi not only significantly outperforms **train-from-scratch** models like SwinIR [36] (+2.11 PSNR) and MambaIR [7] (+1.35 PSNR), but it also comprehensively surpasses **pre-trained foundation model** UniFMIR [6] across both fully fine-tuned (+0.85 PSNR) and efficiently fine-tuned (+3.19 PSNR) configurations. An intriguing finding is that our results indicate Orochi with PEFT leads the list. This outcome is plausible given that the dataset, derived from

Table 3: **Inter-patient Brain Registration Task.** During training, the model goal is to input paired MRI data from distinct patients and output prediction of the registration flow. This flow is applied to the corresponding segmentation data to calculate the dice loss. Thereby, regional deformation can be learned with supervision.

Method	Dice ↑	HD95 ↓	SDlogJ ↓
Dataset: *OASIS* [58]			
Initial	56.10	3.86	—
Lv *et al.* [26]	80.00	1.77	0.08
Siebert *et al.* [27]	81.00	1.63	0.07
Mok *et al.* [28]	82.00	1.67	0.07
PIMed [31]	78.76	1.86	**0.06**
LapIRN [30]	82.18	1.67	0.08
ConvexAdam [29]	81.20	1.71	0.07
Transmorph-B [32]	81.62	1.69	0.12
Transmorph-L [32]	82.22	1.66	0.12
Mambamorph [33]	81.81	1.66	0.09
Orochi (Full)	**83.62**	**1.60**	0.11
Orochi (Normal)	82.52	1.65	0.12
Orochi (Light)	79.61	1.73	**0.06**

Table 4: **CT-MRI Fusion Task.** Volumetric MRI and CT data are sent jointly to the model, reconstructing a single fused result. This fused result would be compared with both MRI and CT input for similarity calculation (e.g. SSIM).

Method	Qabf ↑	Qcv ↓	SSIM ↑
Dataset: *VIFB* [57]			
U2Fusion [14]	0.32	6,580.80	0.41
EMMA [15]	0.29	6,695.80	1.18
ALMFnet [17]	0.29	7,200.50	1.27
MsgFusion [18]	0.19	7,090.40	0.29
MDHU [19]	0.22	7,417.60	1.23
UMF-CMGR [20]	0.25	4,638.70	1.35
SuperFusion [21]	0.28	4,828.90	0.97
MURF [22]	0.33	5,554.60	1.27
IMF [23]	0.27	4,439.60	1.34
PAMRFuse [24]	0.09	5,408.00	0.18
BSAFusion [25]	0.39	4,155.10	1.38
DDFM [16]	0.26	5,981.40	1.31
Orochi (Full)	**0.41**	**2,351.57**	1.39
Orochi (Normal)	0.37	2,519.36	**1.45**
Orochi (Light)	0.34	2,461.41	1.43

isotropic data pairs, comprises fewer than 100 total training patches. Consequently, Full Fine-Tuning or training from scratch is prone to over-fitting. (see Appendix C.3 for extra comparisons)

Image Super-resolution Task We next evaluated the image super-resolution capabilities of Orochi (see Table 2). Early super-resolution models typically rely on CNN-based architectures such as UniRes [10] and SynthSR [12], which are efficient yet often lack sufficient expressiveness and generalization ability. LIIF [13] leverages the power of Implicit Neural Representations (INR) to perform implicit interpolation; however, the high training cost associated with INR limits its adaptability to real-world scenarios. More recent approaches, including InverseSR [11] and VCM [5], based on powerful pre-trained Brain-Latent Diffusion Models (LDM) [59] to overcome these shortcomings. In this setting, Orochi significantly outperforms all the aforementioned architectures. At an 8mm slice thickness, Orochi achieves a PSNR that is 4.01 points higher than InverseSR and 2.76 points higher than VCM. These gains demonstrate that **among pre-trained models, Orochi's pre-training is markedly superior to that of Brain-LDM, both in terms of the pre-training data and purpose.**

Image Registration Task We further evaluated the registration task using the dataset from Learn2Reg [60] (see Table 3). In this task, brain MRI images from different patients (i.e., inter-patients) are registered (see Appendix C.2 for patient-to-atlas brain registration test), and the model's ability to handle subtle deformations is assessed by measuring the similarity of the segmented brain regions after registration (e.g. Dice). Biomedical image registration has evolved from CNN-based [29, 30, 27] to Transformer-based architectures [32, 54], with even linear-complexity models such as Mamba [33] emerging in recent work. In comparison to these methods, **our approach achieves Dice scores that are 2.42 points higher than ConvexAdam, 2.0 points higher than Transmorph, and 1.81 points higher than Mambamorph.**

Image Fusion Task Finally, as illustrated in Table 4, we evaluated Orochi's performance on the image fusion task. Recent trends in this domain have integrated image registration as an auxiliary task to facilitate fusion, as demonstrated by methods such as BSAFusion [25], UMF-CMGR [20], MURF [22], and SuperFusion [21]. Although these models typically exhibit limited registration capabilities (see Appendix C.2), this aligns with our pursuit of developing a versatile, comprehensive model. Compared with the recent advanced model BSAFusion, Orochi outperforms on all evaluated metrics, achieving improvements of +0.02 in Q_{abf}, -1803.53 in Q_{cv}, and +0.07 in SSIM. Combined with our state-of-the-art performance on the registration task, **these results establish Orochi as the first model in this domain to achieve such performance.**

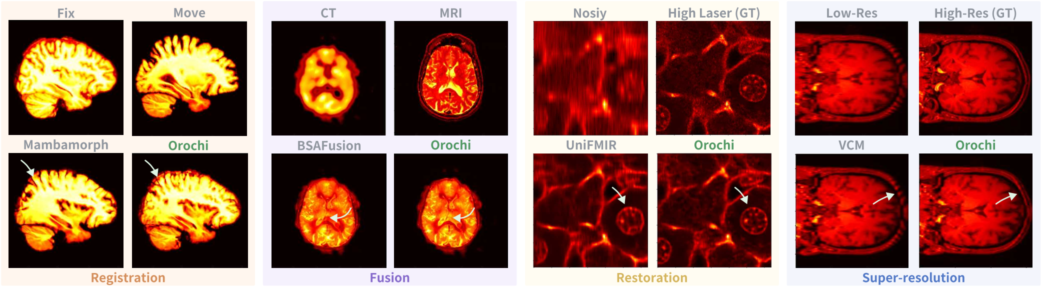


Figure 4: **Visualization Comparison to Recent Advances.** Concise visualization for each task is provided. To enhance comprehension, arrows have been incorporated to facilitate evaluation.

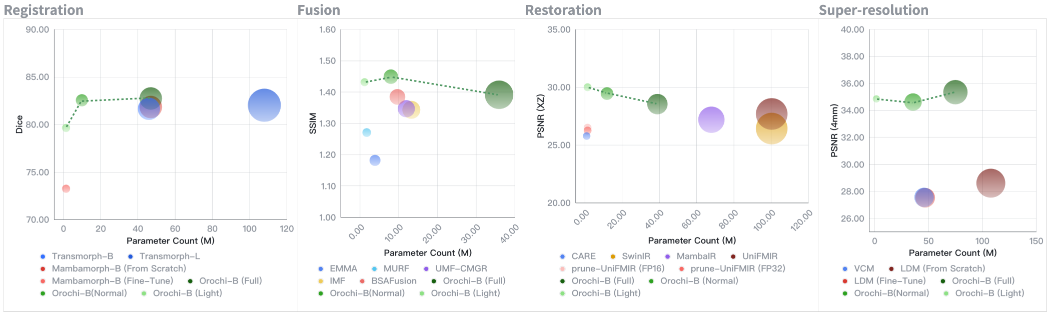


Figure 5: **Fine-Tuning Efficiency V.S Performance.** We present the training parameter efficiency of various models alongside their corresponding results. The three-tier fine-tuning results for Orochi (Full, Normal, Light) are illustrated using a gradient of green colours, from deep to light, and are connected by green dashed lines to indicate the trend. Other baselines are shown in the legend.

Visualizations In Figure 4, we provide qualitative results of Orochi. Specifically, Orochi demonstrates a superior capability in **handling subtle degradations**. (see Appendix C.1 C.2 for more)

Table 5: **Pre-train Strategies V.S Performance.** Datasets and setups remain the same as those experiments in the previous sections.

Strategy	**Registration** (Dice ↑)	**Fusion** (Q_{abf} ↑)	**Restoration** (PSNR ↑)	**Super-Resolution** (PSNR ↑)
MAE [37] (Single Mask)	71.22	0.36	26.67	29.17
I-JEPA [40] (Dual Mask)	69.97	0.39	25.02	28.81
Orochi (TJP)	**83.62**	**0.41**	**29.88**	**33.63**

Ablation Study - Comparison to Other Pre-train Strategies In Table 5, we demonstrate the limitations of relying solely on Masked-image-Modelling (MIM), particularly in registration tasks. Additionally, we observe that the dual-masking approach employed in I-JEPA [40] underperforms compared to Orochi. We hypothesize that this is because chunk masking is more advantageous for high-level tasks rather than the low-level focus of our study.

Ablation Study - Larger ≠ Better, Fine-Tuning Efficiency V.S Performance As shown in Figure 5, the number of trainable parameters is not the decisive factor for downstream tasks—particularly in data-limited scenarios such as biomedical imaging. In many cases, opting for **Parameter-Efficient Fine-Tuning (using only 1–2% of the total parameter count) prevents overfitting and achieves both efficient and effective results.**

5 Conclusion

We introduce **Orochi**, the first versatile biomedical image processor designed for low-level tasks. **To enhance effectiveness,** we propose Random Multi-scale Sampling, which is a scalable way to leverage raw data from a wide range of studies. The extracted data is then processed through our Task-related Joint-embedding Pre-training (TJP), where a unified and robust embedding is learned from various task-related degradations. **For efficiency**, we developed Multi-head Hierarchy Mamba and provide a three-tier fine-tuning framework (Full, Normal, and Light). These design choices ensure high efficiency during pre-training, post-tuning, and test inference. **Our experiments** demonstrate that Orochi exhibits in-domain generalization capability across multiple tasks and achieves state-of-the-art performance compared to specialist models with efficient fine-tuning (less than 5% of total parameters). This suggests that constructing a generalist image processor may lie **more in the diversity of the dataset and the pre-training strategy than in increasing the model size naively.**

6 Acknowledgements

G.D. was supported by the National Natural Science Foundation of China (Grant No. W2442028). S.Z. was supported by the National Science and Technology Major Project (Grant No. 2022ZD0117800). The work of ISAS was supported by the "Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen" and "Der Regierende Bürgermeister von Berlin, Senatskanzlei Wissenschaft und Forschung." The work was further supported by the Bundesministerium für Forschung, Technologie und Raumfahrt, BMFTR under the funding reference 161L0272.

References

[1] Shanghang Zhang, Gaole Dai, Tiejun Huang, and Jianxu Chen. Multimodal large language models for bioimage analysis. *nature methods*, 21(8):1390–1393, 2024.

[2] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[3] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[4] Martin Weigert, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Akanksha Jain, Benjamin Wilhelm, Deborah Schmidt, Coleman Broaddus, Siân Culley, et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nature methods*, 15(12):1090–1097, 2018.

[5] Suhyun Ahn, Wonjung Park, Jihoon Cho, Seunghyuck Park, and Jinah Park. Volumetric conditioning module to control pretrained diffusion models for 3d medical images. *arXiv preprint arXiv:2410.21826*, 2024.

[6] Chenxi Ma, Weimin Tan, Ruian He, and Bo Yan. Pretraining a foundation model for generalizable fluorescence microscopy-based image restoration. *Nature Methods*, 21(8):1558–1567, 2024.

[7] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, pages 222–241. Springer, 2024.

[8] Xuesong Li, Yicong Wu, Yijun Su, Ivan Rey-Suarez, Claudia Matthaeus, Taylor B Updegrove, Zhuang Wei, Lixia Zhang, Hideki Sasaki, Yue Li, et al. Three-dimensional structured illumination microscopy with enhanced axial resolution. *Nature biotechnology*, 41(9):1307–1319, 2023.

[9] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[10] Mikael Brudfors, Yael Balbastre, Parashkev Nachev, and John Ashburner. A tool for super-resolving multimodal clinical mri. *arXiv preprint arXiv:1909.01140*, 2019.

[11] Jueqi Wang, Jacob Levman, Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, M Jorge Cardoso, and Razvan Marinescu. Inversesr: 3d brain mri super-resolution using a latent diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 438–447. Springer, 2023.

[12] Juan Eugenio Iglesias, Benjamin Billot, Yaël Balbastre, Azadeh Tabari, John Conklin, R Gilberto González, Daniel C Alexander, Polina Golland, Brian L Edlow, Bruce Fischl, et al. Joint super-resolution and synthesis of 1 mm isotropic mp-rage volumes from clinical mri exams with scans of different orientation, resolution and contrast. *Neuroimage*, 237:118206, 2021.

[13] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021.

[14] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):502–518, 2020.

[15] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25912–25921, 2024.

[16] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8082–8093, 2023.

[17] Pan Mu, Guanyao Wu, Jinyuan Liu, Yuduo Zhang, Xin Fan, and Risheng Liu. Learning to search a lightweight generalized network for medical image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):5921–5934, 2023.

[18] Jinyu Wen, Feiwei Qin, Jiao Du, Meie Fang, Xinhua Wei, CL Philip Chen, and Ping Li. Msgfusion: Medical semantic guided two-branch network for multimodal brain image fusion. *IEEE Transactions on Multimedia*, 26:944–957, 2023.

[19] Yuchan Jie, Xiaosong Li, Haishu Tan, Fuqiang Zhou, and Gao Wang. Multi-modal medical image fusion via multi-dictionary and truncated huber filtering. *Biomedical Signal Processing and Control*, 88:105671, 2024.

[20] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876*, 2022.

[21] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022.

[22] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):12148–12166, 2023.

[23] Di Wang, Jinyuan Liu, Long Ma, Risheng Liu, and Xin Fan. Improving misaligned multi-modality image fusion with one-stage progressive dense registration. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[24] Yutian Zhong, Jinchuan He, Zhichao Liang, Shuangyang Zhang, Qianjin Feng, Wufan Chen, and Li Qi. Performance of medical image fusion in high-level analysis tasks: A mutual enhancement framework for unaligned pat and mri image fusion. *arXiv preprint arXiv:2407.03992*, 2024.

[25] Huafeng Li, Dayong Su, Qing Cai, and Yafei Zhang. Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion. *arXiv preprint arXiv:2412.08050*, 2024.

[26] Jinxin Lv, Zhiwei Wang, Hongkuan Shi, Haobo Zhang, Sheng Wang, Yilang Wang, and Qiang Li. Joint progressive and coarse-to-fine registration of brain mri via deformation field integration and non-rigid feature fusion. *IEEE Transactions on Medical Imaging*, 41(10):2788–2802, 2022.

[27] Hanna Siebert, Lasse Hansen, and Mattias P. Heinrich. Fast 3d registration with accurate optimisation and little learning for learn2reg 2021, 2021.

[28] Tony CW Mok and Albert CS Chung. Conditional deformable image registration with convolutional neural network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pages 35–45. Springer, 2021.

[29] Hanna Siebert, Christoph Großbröhmer, Lasse Hansen, and Mattias P Heinrich. Convexadam: Self-configuring dual-optimisation-based 3d multitask medical image registration. *IEEE Transactions on Medical Imaging*, 2024.

[30] Tony CW Mok and Albert CS Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 211–221. Springer, 2020.

[31] B Golestani. The learn2reg 2021 miccai grand challenge (pimed team). *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis: MICCAI 2021 Challenges: MIDOG 2021, MOOD 2021, and Learn2Reg 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27–October 1, 2021, Proceedings*, 13166:168, 2022.

[32] 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, 2022.

[33] Tao Guo, Yinuo Wang, and Cai Meng. Mambamorph: a mamba-based backbone with contrastive feature learning for deformable mr-ct registration. *arXiv e-prints*, pages arXiv–2401, 2024.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[35] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[37] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[38] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[40] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[41] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

[42] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[43] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

[44] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021.

[45] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[46] Gaole Dai, Yiming Tang, Chunkai Fan, Qizhe Zhang, Zhi Zhang, Yulu Gan, Chengqing Zeng, Shanghang Zhang, and Tiejun Huang. Discovering long-term effects on parameter efficient fine-tuning. *arXiv preprint arXiv:2409.06706*, 2024.

[47] Caroline A Schneider, Wayne S Rasband, and Kevin W Eliceiri. Nih image to imagej: 25 years of image analysis. *Nature methods*, 9(7):671–675, 2012.

[48] Napari Contributors. napari: a multi-dimensional image viewer for python. *Zenodo https://doi.org/10.5281/zenodo*, 3555620, 2019.

[49] Eleanor Williams, Josh Moore, Simon W Li, Gabriella Rustici, Aleksandra Tarkowska, Anatole Chessel, Simone Leo, Bálint Antal, Richard K Ferguson, Ugis Sarkans, et al. Image data resource: a bioimage data integration and publication platform. *Nature methods*, 14(8):775–781, 2017.

[50] Claire L Walsh, P Tafforeau, WL Wagner, DJ Jafree, A Bellier, C Werlein, MP Kühnel, E Boller, S Walker-Samuel, JL Robertus, et al. Imaging intact human organs with local resolution of cellular structures using hierarchical phase-contrast tomography. *Nature methods*, 18(12):1532–1541, 2021.

[51] Matheus P Viana, Jianxu Chen, Theo A Knijnenburg, Ritvik Vasan, Calysta Yan, Joy E Arakaki, Matte Bailey, Ben Berry, Antoine Borensztejn, Eva M Brown, et al. Integrated intracellular organization and its variations in human ips cells. *Nature*, 613(7943):345–354, 2023.

[52] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[53] Linjing Fang, Fred Monroe, Sammy Weiser Novak, Lyndsey Kirk, Cara R. Schiavon, Seungyoon Blenda Yu, Tong Zhang, Melissa Wu, Kyle Kastner, Alaa Abdel Latif, Zijun Lin, A. Shaw, Yoshiyuki Kubota, John M. Mendenhall, Zhao Zhang, Gulcin Pekkurnaz, Kristen M. Harris, Jeremy Howard, and Uri Manor. Deep learning-based point-scanning super-resolution imaging. *Nature methods*, 18:406 – 416, 2019.

[54] 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.

[55] Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024.

[56] Jingxue Huang, Xiaosong Li, Haishu Tan, and Xiaoqi Cheng. Generative adversarial network for trimodal medical image fusion using primitive relationship reasoning. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[57] D Summers. Harvard whole brain atlas: www. med. harvard. edu/aanlib/home. html. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(3):288–288, 2003.

[58] Elizabeth M Sweeney, Russell T Shinohara, Navid Shiee, Farrah J Mateen, Avni A Chudgar, Jennifer L Cuzzocreo, Peter A Calabresi, Dzung L Pham, Daniel S Reich, and Ciprian M Crainiceanu. Oasis is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in mri. *NeuroImage: clinical*, 2:402–413, 2013.

[59] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.

[60] 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.

[61] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

[62] Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010.

[63] 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.

[64] Mattias P Heinrich, Oskar Maier, and Heinz Handels. Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities. *Visceral Challenge@ ISBI*, 1390:27, 2015.

[65] Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye. Cyclemorph: cycle consistent unsupervised deformable image registration. *Medical image analysis*, 71:102036, 2021.

[66] Huaqi Qiu, Chen Qin, Andreas Schuh, Kerstin Hammernik, and Daniel Rueckert. Learning diffeomorphic and modality-invariant registration using b-splines. In *Medical imaging with deep learning*, 2021.

[67] Junyu Chen, Yufan He, Eric C Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*, 2021.

[68] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

[69] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 171–180. Springer, 2021.

[70] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: volumetric medical image segmentation via a 3d transformer. *IEEE transactions on image processing*, 32:4036–4045, 2023.

[71] Chang Qiao, Di Li, Yuting Guo, Chong Liu, Tao Jiang, Qionghai Dai, and Dong Li. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nature methods*, 18(2):194–202, 2021.

[72] Bin Xia, Yucheng Hang, Yapeng Tian, Wenming Yang, Qingmin Liao, and Jie Zhou. Efficient non-local contrastive attention for image super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2759–2767, 2022.

[73] Yu Kang T Xu, Austin R Graves, Gabrielle I Coste, Richard L Huganir, Dwight E Bergles, Adam S Charles, and Jeremias Sulam. Cross-modality supervised image restoration enables nanoscale tracking of synaptic plasticity in living mice. *Nature Methods*, 20(6):935–944, 2023.

[74] Chang Qiao, Di Li, Yuting Guo, Chong Liu, Tao Jiang, Qionghai Dai, and Dong Li. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nature methods*, 18(2):194–202, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions are well justified with comprehensive theoretical and experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the full set of assumptions and a complete (and correct) proof

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a condensed implementation in the experiment section and a detailed description in the Appendix, with code submitted in the supplemental materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: A Readme.md file is attached along with the code submitted in supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the implementation details are included in the Appendix section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our experiments are conducted with a set random seed 42.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We make sure to preserve anonymity.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the assets are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper currently does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Experiment Setups

A.1 Baselines

Registration

- **Lv et al.** [26]: Uses a dual-encoder U-Net for coarse-to-fine registration.
- **Siebert et al.** [27]: Proposes a fast 3D registration approach.
- **Mok et al.** [28]: Employs conditional deformable convolutions.
- **PIMed** [31]: From the Learn2Reg challenge.
- **LapIRN** [30]: Uses a Laplacian pyramid network for large deformations.
- **ConvexAdam** [29]: Adopts a dual-optimization strategy.
- **TransMorph** [32]: Based on a Transformer architecture for capturing long-range correspondences.
- **MambaMorph** [33]: Utilizes mamba blocks for efficient long-range dependency modeling.

Fuison

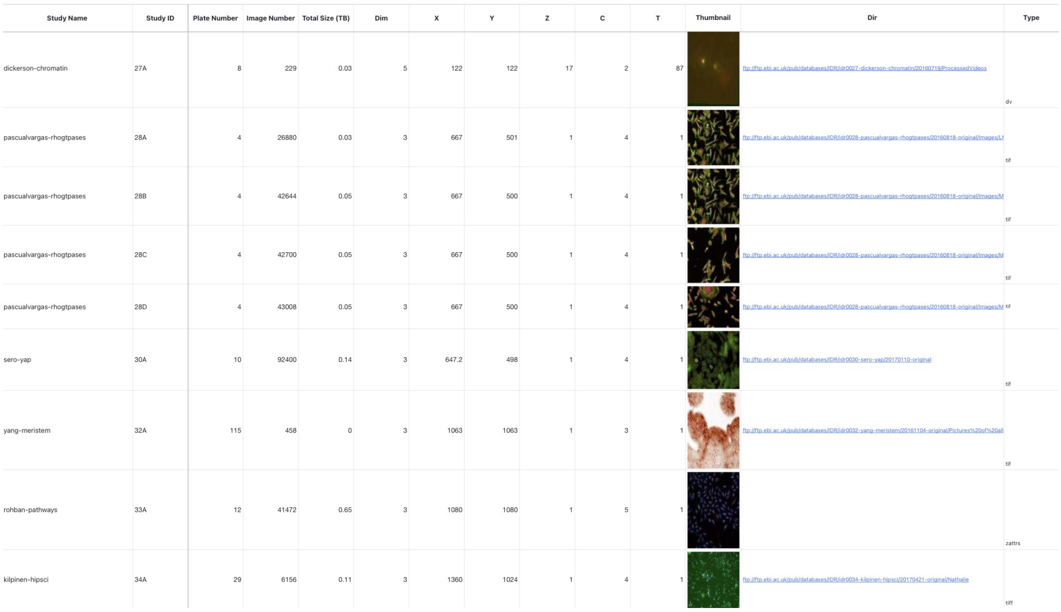
- **U2Fusion** [14]: Provides a unified unsupervised fusion approach.
- **EMMA** [15]: Employs equivariant learning for fusion.
- **ALMFNet** [17]: Searches for a lightweight generalized fusion network.
- **MsgFusion** [18]: Uses a semantic-guided two-branch network.
- **MDHU** [19]: Uses multi-dictionary learning with truncated Huber filtering.
- **UMF-CMGR** [20]: Adopts cross-modality generation and registration.
- **SuperFusion** [21]: Combines registration and fusion with semantic awareness.
- **MURF** [22]: Reinforces multi-modal registration and fusion mutually.
- **IMF** [23]: Improves fusion with a progressive dense registration strategy.
- **PAMRFuse** [24]: Focuses on feature alignment.
- **BSAFusion** [25]: Adopts bidirectional stepwise feature alignment.
- **DDFM** [16]: Utilizes a denoising diffusion model for fusion.

Super-Resolution

- **Cubic**: Bicubic interpolation as a traditional baseline.
- **UniRes** [10]: Designed for super-resolving multimodal clinical MRI.
- **SynthSR** [12]: Performs joint super-resolution and synthesis.
- **LIIF** [13]: Learns continuous image representations for implicit interpolation.
- **InverseSR** [11]: Uses a latent diffusion model for 3D brain MRI super-resolution.
- **VCM** [5]: Applies a volumetric conditioning module.

Restoration

- **Li et al.** [8]: Improves axial resolution.
- **CARE** [4]: Uses a content-aware network for fluorescence microscopy image restoration.
- **SwinIR** [9]: Employs a Swin-Transformer for efficient image restoration.
- **MambaIR** [7]: Utilizes mamba blocks for modeling long-range dependencies.
- **UniFMIR** [6]: Fine-tunes a pre-trained foundation model for generalizable fluorescence microscopy-based restoration (with pruned FP16/FP32 variants).



Study Name	Study ID	Plate Number	Image Number	Total Size (TB)	Dim	X	Y	Z	C	T	Thumbnail	Dir	Type
dickerson-chromatin	27A	8	229	0.03	5	122	122	17	2	87		ftp://ftp.ebi.ac.uk/pub/databases/IDR/idr0027-dickerson-chromatin/20160719/ProcessedVideos	dv
pascualvargas-rhogtpases	28A	4	26880	0.03	3	667	501	1	4	1		ftp://ftp.ebi.ac.uk/pub/databases/IDR/idr0028-pascualvargas-rhogtpases/20160818-original/Images/L	tif
pascualvargas-rhogtpases	28B	4	42644	0.05	3	667	500	1	4	1		ftp://ftp.ebi.ac.uk/pub/databases/IDR/idr0028-pascualvargas-rhogtpases/20160818-original/Images/M	tif
pascualvargas-rhogtpases	28C	4	42700	0.05	3	667	500	1	4	1		ftp://ftp.ebi.ac.uk/pub/databases/IDR/idr0028-pascualvargas-rhogtpases/20160818-original/Images/M	tif
pascualvargas-rhogtpases	28D	4	43008	0.05	3	667	500	1	4	1		ftp://ftp.ebi.ac.uk/pub/databases/IDR/idr0028-pascualvargas-rhogtpases/20160818-original/Images/M	tif
sero-yap	30A	10	92400	0.14	3	647.2	498	1	4	1		ftp://ftp.ebi.ac.uk/pub/databases/IDR/idr0030-sero-yap/20170110-original	tif
yang-meristem	32A	115	458	0	3	1063	1063	1	3	1		ftp://ftp.ebi.ac.uk/pub/databases/IDR/idr0032-yang-meristem/20161104-original/Pictures%20of%20all	tif
rohban-pathways	33A	12	41472	0.65	3	1080	1080	1	5	1			zattrs
kilpinen-hipsci	34A	29	6156	0.11	3	1360	1024	1	4	1		ftp://ftp.ebi.ac.uk/pub/databases/IDR/idr0034-kilpinen-hipsci/20170421-original/Nathalie	tiff

Figure 6: **Preview of the Metadata List of Studies**

A.2 Datasets

Pre-train

- A combined multi-modal biomedical image dataset aggregated from over 100 public studies, encompassing various imaging modalities and degradation types [51, 50, 49]. In Figure 6, we provide a preview of the metadata of the studies we covered (Excel Form would be included in the Zip file). Since our RMS method is highly scalable, we plan to further update this list in the future and explore the borderline.

Registration

- The OASIS brain MRI dataset from the Learn2Reg 2021 challenge, used to evaluate the overlap of segmented regions and the smoothness of the deformation fields [60, 58].

Fuison

- A CT–MRI paired fusion dataset (VIFB), which assesses the integration of complementary information across modalities [57].

Super-Resolution

- The Harvard Whole Brain Atlas (HBA), providing high-quality MRI images for evaluating low-resolution image reconstruction [57].

Restoration

- The CARE microscopy image dataset, used to evaluate the enhancement of low signal-to-noise ratio fluorescence microscopy images [4].

A.3 Metrics

Registration

- **Dice similarity coefficient:** Computed as

$$Dice = \frac{2|A \cap B|}{|A| + |B|}$$

which measures the overlap between the segmented regions.

- **95th percentile Hausdorff Distance (HD95):** Defined as the 95th percentile of the distances between boundary points of the segmented regions.
- **Standard deviation of the log-Jacobian determinant (SDlogJ):** Calculated as the standard deviation of $\log(\det(J))$, where J is the Jacobian matrix of the deformation field. This metric reflects the smoothness of the deformation field [60].

Fuison

- $Q_{\mathbf{AB/F}}$ **($\mathbf{Q_{abf}}$):** Measures the quality of the fusion by evaluating the consistency between the fused image and the input modalities.
- $Q_{\mathbf{CV}}$ **($\mathbf{Q_{cv}}$):** Assesses the contrast consistency across the fused image.
- **Structural Similarity Index (SSIM):** Computed based on comparisons of luminance, contrast, and structure between 2 source images [25].

Super-Resolution

- **Peak Signal-to-Noise Ratio (PSNR):** Calculated as

$$PSNR = 10\log_{10}\left(\frac{\mathrm{MAX}_I^2}{MSE}\right)$$

  where MAX_I is the maximum possible pixel value and MSE is the mean squared error between the reconstructed and reference images.
- **SSIM:** Evaluates perceptual similarity between the super-resolved and reference images [9].

Restoration

- **PSNR:** As above, it measures the pixel-level fidelity between the restored image and the high-quality reference.
- **SSIM:** Measures the structural similarity between the restored and reference images [4].

A.4 Code-base

Pre-train

- Adapted from public GitHub implementations of the Swin-Transformer and Transmorph. **Swin-Transformer:** `https://github.com/microsoft/Swin-Transformer` [36]; **Transmorph:** `https://github.com/junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration` [32].

Registration

- Implemented based on the Transmorph GitHub code. **Link:** `https://github.com/junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration` [32].

Fuison

- Built with reference to the BSAFusion GitHub code. **Link:** `https://github.com/slrl123/BSAFusion` [25].

Super-Resolution

- Implemented based on GitHub codes of InverseSR and VCM. **InverseSR:** `https://github.com/BioMedAI-UCSC/InverseSR` [11]; **VCM:** `https://github.com/Ahn-Ssu/VCM` [5].

Restoration

- Based on the UniFMIR GitHub implementation. **Link:** `https://github.com/cxm12/UNiFMIR` [6].

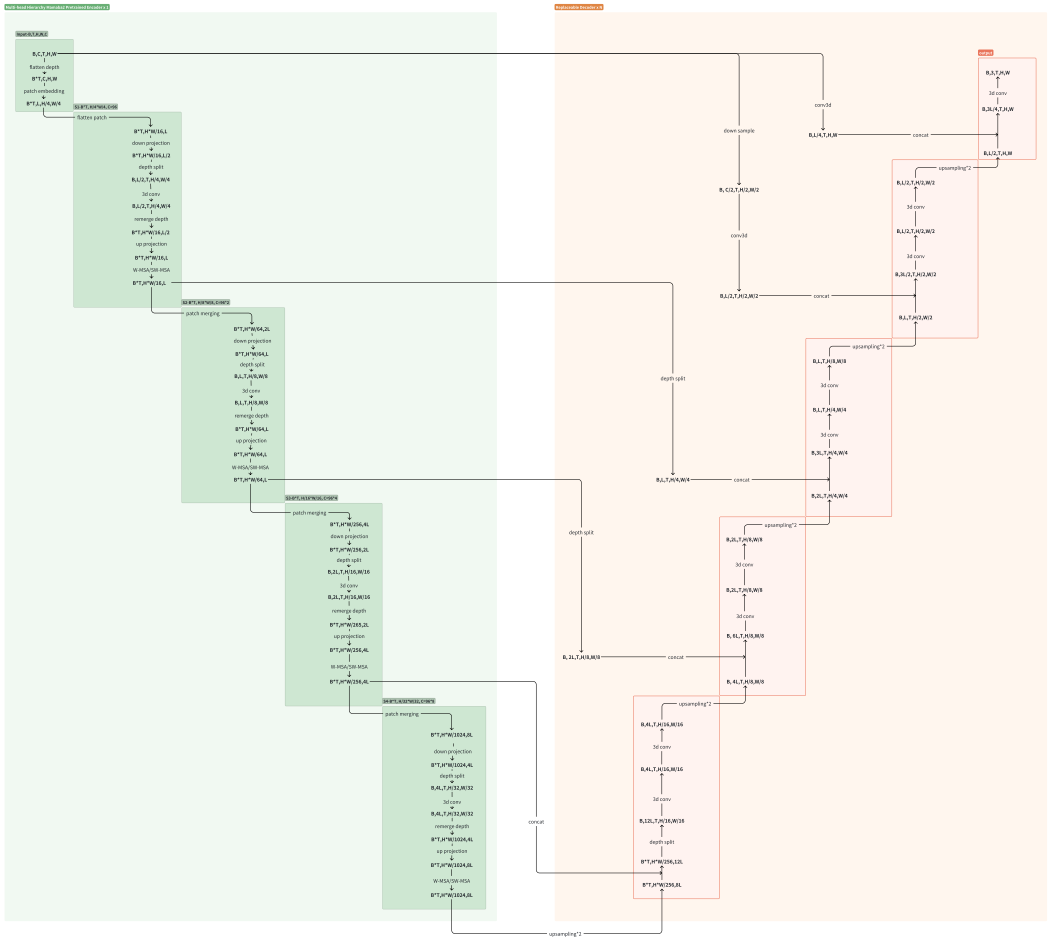


Figure 7: **Model architecture of Multi-head Hierarchy Mamba (MHM).** The model contains a unified hierarchical Mamba encoder with a replaceable decoder (e.g. regular convolution head or depth-wise separable convolution head) for tuning

B Experiment Configurations

Multi-head Hierarchy Mamba & Three-Tier Fine-Tuning Framework Figure 7, presents a comprehensive diagram of Orochi's backbone architecture. Post-tuning, the interchangeable decoder can be replaced as required. We evaluated Orochi's performance using our Three-Tier Fine-Tuning Framework, which includes full fine-tuning (Full, 100% parameters), regular convolution head with the encoder frozen (Normal, 10-30% parameters), and depth-wise separable convolution head [45] with the encoder frozen (Light, less than 5% parameters). The optimal results were achieved across all three tiers, underscoring the significance of selecting an appropriate tuning method based on specific requirements.

Pre-train We pre-trained Orochi-B (3D version) with the configuration listed in Table 6. The 2D version has a similar configuration, with slight differences on some setups (e.g. batch size). We have 2 two sets of pre-training devices. The A800 80Gx8 device is used for local pre-train and the H100 40Gx8 device is for streaming pre-train.

Fine-tuning We followed the same setups as our code base for each task (see Section **??**), including the tuning resolution, epoch number, optimizer configurations and loss designs. The device we use for fine-tuning is NVIDIA 4090 24Gx4

Table 6: **Pretraining Configuration Parameters for Orochi-B**

Category	**Parameter**	**Value / Range and Description**
Model / Encoder	`img_size`	(32, 224, 224)
	`patch_size`	4
	`pat_merg_rf`	2
	`in_chans`	2
	`embed_dim`	128
	`depths`	(4, 4, 4, 4)
	`drop_path_rate`	0.2
	`if_convskip`	True
	`out_indices`	(0, 1, 2, 3)
Mamba	`ssm_cfg`	None
	`norm_epsilon`	1×10^{-5}
	`initializer_cfg`	None
	`fused_add_norm`	True
	`rms_norm`	True
	`residual_in_fp32`	True
	`patch_norm`	True
	`use_checkpoint`	False
Decoder	`decoder_bn`	False
	`decoder_depthseparable`	False
	`decoder_head_chan`	64
Training	`batch_size`	12
	`lr`	0.0005
	`weight_decay`	0.01
	`warmup_ratio`	0.1
	`warmup_start_factor`	0.01
	`max_epoch`	50
	Optimizer / Scheduler	AdamW; WarmupCosine with cycles=0.5
Deformation & Augmentation	Registration Flow Scaling	$\tanh(\cdot)\times 0.6$ (applied to the deformation field)
	Registration Gaussian Sigma Range	[1.5, 3.5]
	Perlin Noise Octaves	4
	Perlin Noise Persistence	0.5
	Mask Ratio	0.5
	Downsampling Scale Factor Range	[0.25, 0.75]
	Downsampling Noise Level Range	[0.01, 0.1]
	Downsampling Gaussian Sigma Range	[0.25, 1.0]
	Gaussian Noise Level Range	[0.075, 0.15]
	Salt vs. Pepper Ratio	0.5
	Salt & Pepper Noise Amount Range	[0.01, 0.05]
	Grid Image Parameters	Grid spacing = 4, Line width = 1

C Extra Results

C.1 Zero-shot Processing on Biomedical Images

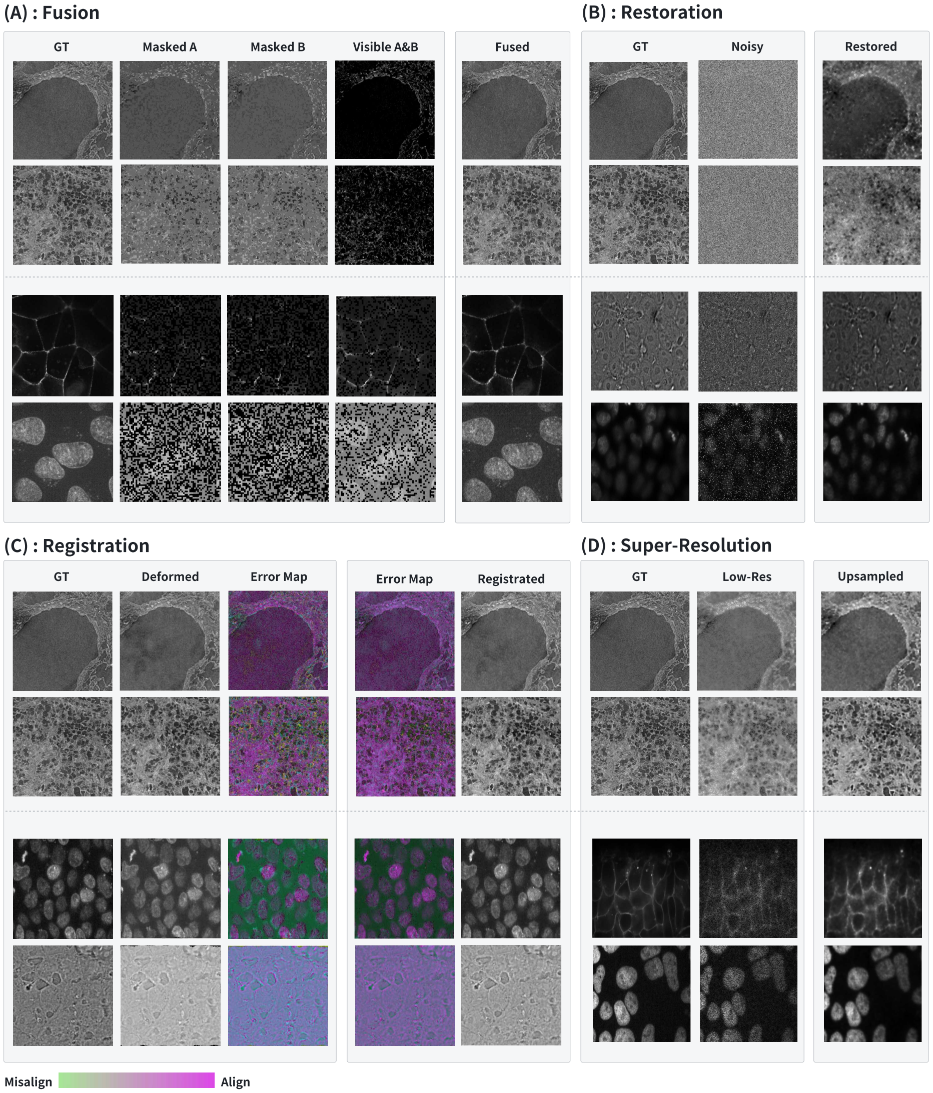


Figure 8: **Visualization of Orochi Zero-shot Processing to Different Degradation**. (A)-(D) panel shows the ability on four tasks respectively, with the dotted line separating medical images (top) and microscopy images (bottom).

In Figure 8, we present additional results demonstrating Orochi's zero-shot performance on both microscopy and medical images. Notably, Orochi yields satisfactory outcomes even when faced with extremely severe degradation, as exemplified in panel (B), row 2, and panel (C), rows 2 and 3.

C.2 Registration & Fuison

Fusion Model on Registration Task In Figure 9, we demonstrate that, despite the recent trend of pre-registration before fusion, these methods remain predominantly fusion-oriented and are not well-suited for addressing real-world registration tasks within the medical image registration community. However, Orochi represents a significant advancement in versatility, as it is designed not only for this specific scenario but also to achieve superior performance across all registration tasks.

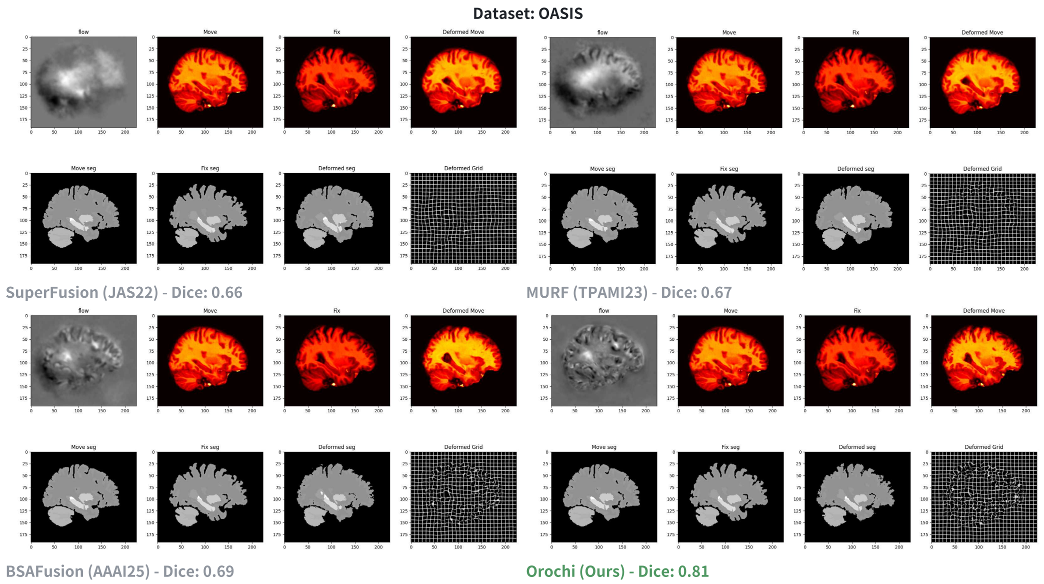

Dataset: OASIS
flow
Move
Fix
Deformed Move
Move seg
Fix seg
Deformed seg
Deformed Grid
SuperFusion (JAS22) - Dice: 0.66
MURF (TPAMI23) - Dice: 0.67
BSAFusion (AAAI25) - Dice: 0.69
Orochi (Ours) - Dice: 0.81

Figure 9: **Performance of Fusion-oriented methods on specialized registration benchmark.** We provide the inputs (Move, Fix), output (flow), and evaluation data (Move/Fix seg and Deformation Grid) for each visualization.

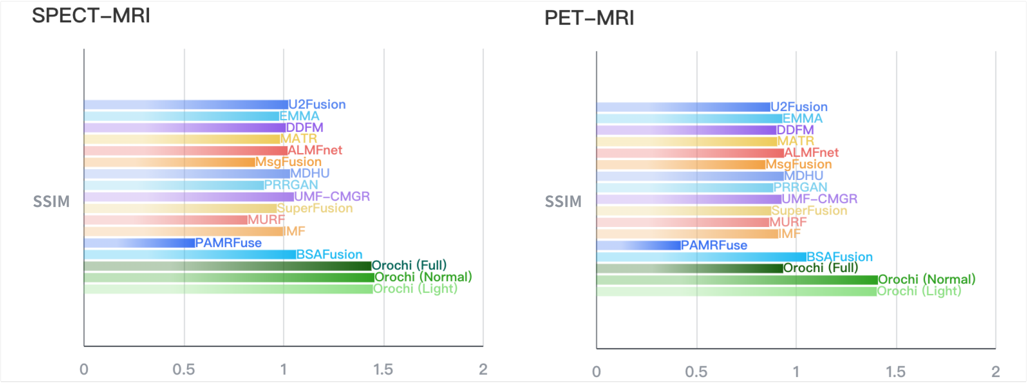


Figure 10: **Fusion Comparison with Brain SPECT&MRI / PET&MRI Data.**

Patient to Atlas Brain Image Registration The regional deformation is learned unsupervised in Table 7. Only the raw image of the atlas and the patient's brain would be used for loss calculation while training. Then we evaluate the dice score between the segmentation maps of these two brains. Since Orochi is pre-trained in this unsupervised fashion, it shows excellent adaptation to this task, similar to the case with supervision.

SPECT-MRI & PET-MRI Image Fusion In Figure 10, we performed comparative evaluations using state-of-the-art fusion techniques on two additional Harvard Whole Brain datasets obtained from `https://www.med.harvard.edu/aanlib/`. These datasets specifically focus on the fusion of SPECT and PET imaging with MRI. The results demonstrate that Orochi outperforms recent advancements such as BSAFusion and maintains superior efficiency.

Table 7: **Patient to Atlas Brain Registration Task.** During the training phase, the model aims to input paired MRI data from both the standard brain atlas and patient scans, to output a predicted registration flow. This flow is subsequently applied to the atlas data to compute the similarity between the registered atlas and the patient's scan. During the testing phase, the predicted flow is applied to the atlas brain segmentation map, and the Dice coefficient is evaluated against the patient's brain segmentation map.

Method	**Dice** ↑	% of $\|J_\Phi\| \leq 0$
Dataset: ***IXI*** **[32]**		
Affine	0.386 ± 0.195	—
SyN [61]	0.645 ± 0.152	≤ 0.0001
NiftyReg [62]	0.645 ± 0.167	0.020 ± 0.046
LDDMM [63]	0.680 ± 0.135	≤ 0.0001
deedsBCV [64]	0.733 ± 0.126	0.147 ± 0.050
VoxelMorph-1 [54]	0.729 ± 0.129	1.590 ± 0.339
VoxelMorph-2 [54]	0.732 ± 0.123	1.522 ± 0.336
VoxelMorph-diff [54]	0.580 ± 0.165	≤ 0.0001
CycleMorph [65]	0.737 ± 0.123	1.719 ± 0.382
MIDIR [66]	0.742 ± 0.128	≤ 0.0001
ViT-V-Net [67]	0.734 ± 0.124	1.609 ± 0.319
PVT [68]	0.727 ± 0.128	1.858 ± 0.314
CoTr [69]	0.735 ± 0.135	1.292 ± 0.342
nnFormer [70]	0.747 ± 0.135	1.595 ± 0.358
TransMorph-Bayes [32]	0.753 ± 0.123	1.560 ± 0.333
TransMorph-diff [32]	0.594 ± 0.163	≤ 0.0001
TransMorph-bspl [32]	0.761 ± 0.122	≤ 0.0001
TransMorph [32]	0.754 ± 0.124	1.579 ± 0.328
Orochi (Full)	**0.770 ± 0.120**	1.592 ± 0.334
Orochi (Normal)	0.765 ± 0.121	1.571 ± 0.323
Orochi (Light)	0.752 ± 0.126	1.499 ± 0.301

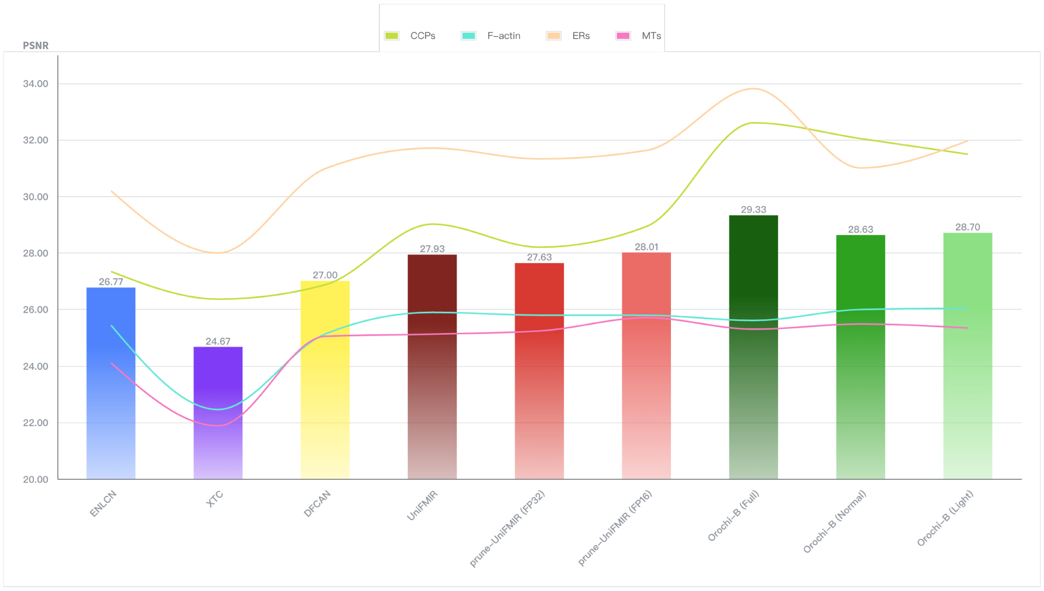


Figure 11: **Stress Test on BioSR Benchmark.** We trained Orochi on four subsets concurrently, thereby reducing the cost associated with hyperparameter searching. The results were compared against baselines that involved separate hyperparameter searches for each subset. The bar chart illustrates the average PSNR values, while the lines, colour-coded according to the legends, indicate the performance metrics for each respective subset.

C.3 Super-Resolution & Restoration

Stress test on joint multi-modal data image repairing In this additional validation, we aim to evaluate Orochi's performance under stress using an extended benchmark. The BioSR [71] benchmark comprises four distinct categories of microscopy image pairs (x2 low/high imaging quality), captured by a multimodal structured illumination microscopy (SIM) system, encompassing Clathrin-Coated Pits (CCPs), Endoplasmic Reticula (ERs), Microtubules (MTs), and F-actin Filaments. Specifically, Orochi was trained on all four datasets concurrently, whereas the baseline [72, 73, 74, 6] models were trained separately on each dataset. This deliberate approach highlights Orochi's capability in resource-constrained environments, where conducting hyperparameter searches for each subset is not feasible. As illustrated in Figure 11, despite the training constraints imposed on Orochi, an absolute improvement is still observed, further demonstrating its capability and efficiency.

D Limitations

Two limitations in our paper remain unaddressed at present. First, due to constraints on computational resources and group size, we were unable to further investigate the scaling law of our method during pre-training. This limitation also indicates that our focus was restricted to low-level tasks as presented in the paper. However, we firmly believe that a unified model for life sciences, capable of excelling in both high-level understanding tasks and low-level generation tasks, will emerge in the future. This is also an emerging trend that has already demonstrated progress in general applications.