Pattern Recognition 172 (2026) 112736

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Iterative optimal transport for multimodal image registration ⋆

Mengyu Li [a], Cheng Meng [b,*], Xiaodan Fan [c]

[a] Department of Statistics and Data Science, Tsinghua University, Beijing, China
[b] Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China
[c] Department of Statistics and Data Science, The Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

Keywords:
Multimodal image registration
Optimal transport
Polynomial transformation
Alternating minimization
Medical imaging

ABSTRACT

The process of aligning images from different modalities, a.k.a. multimodal image registration, is a crucial task in various fields such as medical imaging, remote sensing, and computer vision. Traditional matching methods encounter difficulties in handling nonlinear appearance distortions and partial overlaps across modalities. In this paper, we propose a novel and robust multimodal image registration method, referred to as Iterative Optimal Transport (IOT), that formulates registration as a sequence of optimal transport problems. Specifically, by representing images with edge points, we propose a regularized unbalanced optimal transport criterion that robustly aligns structural information across modalities. Unlike prior approaches based on one-shot OT matching, IOT alternates between computing the transport plan and refining the estimated transformation, allowing for progressively improved alignment with theoretical convergence guarantees. Evaluations of three different types of multimodal image datasets, including brain and retina images, demonstrate the superior performance of IOT over state-of-the-art competitors in various scenarios. These results show the effectiveness and robustness of IOT in multimodal image registration.

1. Introduction

Integrating multimodal images can provide complementary information and enhance the precision of decision-making processes. For instance, in the medical field, the combination of anatomical imaging (e.g., computed tomography (CT) and magnetic resonance imaging (MRI)) and functional imaging (e.g., functional MRI (fMRI) and single-photon emission CT (SPECT)) offers a comprehensive view of body structures and functions, critical for accurate tumor contouring in radiotherapy. In computer vision, visible and infrared images provide distinct perspectives by capturing reflected light and thermal radiation, respectively. Combining these two image types is beneficial in wide applications such as fever screening and building inspections [1]. However, effective use of multimodal information requires precise spatial alignment of images of the same or similar scenes from distinct modalities. This problem is known as multimodal image registration, which serves as a fundamental prerequisite for downstream analyses, including image fusion, object detection, and video tracking, among others [2,3].

Despite its importance, multimodal image registration poses significant challenges. Beyond addressing typical geometrical deformations in general image matching problems, it is necessary to address intrinsic differences in imaging mechanisms across modalities. These differences often lead to large nonlinear appearance distortions [4,5], such as variations in resolution and texture between image pairs, adding layers of complexity to the task.

Existing registration techniques generally fall into two categories: intensity-based and feature-based [3,5]. Intensity-based approaches, focusing on maximizing image similarity metrics, are limited to small initial registration errors and often fail under serious geometrical deformations [5,6]. Feature-based methods, which detect and match handcrafted or learnable features, offer greater resilience to image deformations and noise, but struggle with nonlinear appearance distortions inherent in multimodal images [4]. Recent advanced deep learning techniques, capable of learning features automatically, show promise in overcoming these challenges [2,7]. Nevertheless, most deep learning approaches require large amounts of pre-aligned images or labeled landmarks for training [8], which may be inaccessible in practice. Therefore, there is still an urgent need for more efficient and effective tools for multimodal image registration.

In this paper, we bridge this gap by developing a novel multimodal image registration method based on optimal transport (OT) [9]. The OT theory, originated by Gaspard Monge in the 18th century and later

⋆ This work is supported by the Beijing Municipal Natural Science Foundation (Grant No. 1232019), the Renmin University of China Research Fund Program for Young Scholars, and a grant from the Research Grants Council (GRF 14306324) of the Hong Kong SAR, China.

* Corresponding author.
E-mail addresses: mengyuli@tsinghua.edu.cn (M. Li), chengmeng@ruc.edu.cn (C. Meng), xfan@cuhk.edu.hk (X. Fan).

https://doi.org/10.1016/j.patcog.2025.112736
Received 11 June 2025; Received in revised form 6 October 2025; Accepted 11 November 2025
Available online 15 November 2025

developed by Leonid Kantorovich, is a rich mathematical framework that aims to move one distribution of mass (or probability measure) to another with minimum effort. Due to its capability of providing both a valid metric and explicit correspondences for distributions, OT has emerged as a powerful tool in various fields, including machine learning, computer vision, statistics, and biomedical research, among others [10]. In particular, OT theory naturally quantifies uncertainty and deformation with awareness of the underlying geometry, making it especially effective in a variety of imaging applications [11,12]. Moreover, compared with deep learning-based approaches, OT is training-free and can be directly applied to unseen multimodal pairs without supervision, offering a practical alternative when annotated datasets are scarce.

Despite the advantages, existing OT-based image or point registration methods typically rely on one-shot matching [11,13,14] and assume full correspondence between features [13,15,16], and such approaches may fail in multimodal settings. In contrast, we formulate registration as an iterative optimization of transport and transformation. This novel perspective introduces two key benefits: (i) flexibility, enabling progressive improvement of alignment; and (ii) robustness, by incorporating regularized unbalanced OT to handle partial matches, outliers, and modality-induced inconsistencies more effectively.

Contributions. Our major contributions are four-fold.

1. **An iterative OT-based framework for multimodal registration.** We introduce the first framework that models multimodal image registration as an iterative sequence of unbalanced optimal transport (UOT) problems over edge point sets. Unlike classical OT, our unbalanced variant is capable of handling partial alignment, essential for handling modality-specific structures and missing regions. Unlike one-shot OT methods, our iterative approach progressively improves alignment quality. The transformation is modeled in a polynomial space, allowing the method to capture both affine and non-rigid deformations.
2. **A theoretically grounded and efficient algorithm.** We show that the proposed joint optimization problem is biconvex with respect to the transport plan and the transformation parameters. Using this biconvexity, we develop an alternating minimization algorithm with convergence guarantees to solve the problem efficiently.
3. **Comprehensive evaluations in diverse modalities.** We conduct comprehensive experiments using various types of multimodal images, including brain MRI images with varying weights (i.e., T1 and T2), brain images from different medical imaging techniques (i.e., CT and SPECT), and retina images of different angiographies. To our knowledge, our proposed IOT framework is the first to exhibit leading performance across all these types of multimodal images, marking a significant breakthrough in general multimodal image registration.
4. **Complementary and training-free design.** The proposed method is complementary to existing modality-agnostic features, robust similarity measures, and learning-based frameworks. It can integrate modality-robust features or similarities via a hybrid cost formulation. Moreover, unlike deep learning methods that require large paired datasets, IOT is fully training-free and can be directly applied to unseen multimodal pairs, making it particularly valuable when annotated data are scarce.

The remainder of this paper is organized as follows. Section 2 provides an overview of existing image registration methods and the optimal transport theory. Section 3 details the formulation, theoretical properties, and optimization algorithm of the proposed IOT method. In Section 4, we evaluate the performance of our IOT approach through various multimodal images. Supplementary materials include feature-augmented extension, technical proofs, additional implementation details, and extended experiments covering unimodal bechmarks and transformation regularity analyses. The implementation code for the proposed method is available at the following link: https://github.com/Mengyu8042/IOT.

2. Background

We briefly review existing work for image registration, with a focus on intensity-based and feature-based paradigms, followed by the applications of optimal transport theory in this domain and their limitations.

To begin with, we summarize the notation used throughout the paper. We adopt the standard convention of using uppercase boldface letters for matrices, lowercase boldface letters for vectors, and regular font for scalars. Specifically, x_i denotes the ith element of a vector $\boldsymbol{x} \in \mathbb{R}^n$, and X_{ij} represents the (i,j)th element of a matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$. For matrices $\mathbf{X}$ and $\mathbf{Y}$ of the same dimension, their Frobenius inner product is denoted as $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i,j} X_{ij} Y_{ij}$. The Euclidean norm of a vector $\boldsymbol{x}$ is denoted by $\|\boldsymbol{x}\|$.

2.1. Image registration

Image registration seeks a spatial transformation that best aligns one image with the other, traditionally referred to as the “moving" and “fixed" images. Despite this terminology suggesting an asymmetric role, the designation of which image is moving or fixed is flexible and user-defined, catering to specific needs. Existing registration pipelines are typically classified into intensity-based and feature-based [3,5].

Intensity-based methods. These methods deal directly with the intensity values of entire images without requiring explicit feature extraction [17,18]. Given a similarity metric, such as (normalized) cross-correlation or mutual information, along with a transformation model and an optimization method, intensity-based methods maximize the similarity metric between the warped moving image and the fixed image to estimate the transformation parameters. This pipeline can achieve high accuracy if the initial misalignment between images is small. However, it faces two main limitations. First, the similarity metrics may not be linearly related to the accuracy of image registration and are heavily influenced by the size of the overlapping area and the appearance differences between modalities [5]. Second, almost all intensity-based methods are limited to a small range of initial registration errors, but typically fail in cases of severe image deformations [6]. Such challenges are further illustrated through the experimental results presented in Section 4.

Feature-based methods. The feature-based pipeline can be more effective in the face of geometrical deformations [5]. These approaches extract features from images and reduce the task to feature matching, aiming to find the underlying spatial transformation between two sets of extracted features. Traditional feature-based methods, such as the scale-invariant feature transform (SIFT) [19] and histograms of oriented gradient (HOG) [20], rely on appearance attributes like colors, textures, and gradient histograms for feature detection and description. However, such appearance features often no longer match across different modalities. Instead, features representing salient structures, such as corners and edges, are largely preserved within multimodal images and thus preferred for capturing common information [4,21]. Recently, deep neural networks (e.g., convolutional neural networks, Siamese networks, and generative adversarial networks) have been employed to automatically learn features rather than manual design. These techniques typically require labeled training data and are beyond the scope of this paper. We refer to [22] for a comprehensive overview of this line of work.

2.2. Optimal transport for image registration

Consider two sets of points $\{\boldsymbol{x}_i\}_{i=1}^n, \{\boldsymbol{y}_j\}_{j=1}^m \subset \mathbb{R}^d$, each associated with histograms $\boldsymbol{a} \in \Delta^{n-1}$ and $\boldsymbol{b} \in \Delta^{m-1}$, respectively, where $\Delta^{n-1} = \{\boldsymbol{p} \in \mathbb{R}_+^n : \sum_i p_i = 1\}$ represents the $(n-1)$-dimensional standard simplex. Then, a pair of discrete probability measures is defined as

$$\mu = \sum_{i=1}^{n} a_i \delta_{\boldsymbol{x}_i} \quad \text{and} \quad \nu = \sum_{j=1}^{m} b_j \delta_{\boldsymbol{y}_j},$$

where δ_z denotes the Dirac delta function spiking at z. The goal of optimal transport (OT) is to find the most efficient way to move the masses $\boldsymbol{a}$ and $\boldsymbol{b}$ to each other, according to certain ground costs between the support points, $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{y}_j\}$. The modern Kantorovich formulation of OT takes the form

$$\min_{\mathbf{T}\in\Pi(\boldsymbol{a},\boldsymbol{b})} \langle \mathbf{C}, \mathbf{T}\rangle := \sum_{i,j} c(\boldsymbol{x}_i, \boldsymbol{y}_j) T_{ij}, \tag{1}$$

where $\Pi(\boldsymbol{a},\boldsymbol{b}) := \{\mathbf{T} \in \mathbb{R}_+^{n\times m} : \mathbf{T}\mathbf{1}_m = \boldsymbol{a}, \mathbf{T}^\top \mathbf{1}_n = \boldsymbol{b}\}$ is the set of admissible transportation plans, i.e., all joint probability distributions with marginals $\boldsymbol{a}$ and $\boldsymbol{b}$; T_{ij} represents the amount of mass transferred from $\boldsymbol{x}_i$ to $\boldsymbol{y}_j$; and $\mathbf{C} \in \mathbb{R}_+^{n\times m}$ is a cost matrix determined by the cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, where $C_{ij} = c(\boldsymbol{x}_i, \boldsymbol{y}_j)$ defines the pairwise cost of moving one unit of mass from $\boldsymbol{x}_i$ to $\boldsymbol{y}_j$. When $c(\boldsymbol{x}_i, \boldsymbol{y}_j) = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^p$ ($p \in \mathbb{Z}_+$), the optimal objective value of the problem (1) is known as the p-Wasserstein distance, which has been widely used to quantify the discrepancy between distributions. The solution to (1) is called the optimal transport plan, achieving the minimal total cost of transportation.

During the past two decades, by leveraging the correspondences established through the OT plan and the similarity metric provided by the Wasserstein distance, OT-based image registration methods have emerged and evolved, encompassing both intensity-based [15,16] and feature-based approaches [11–14]. For example, Haker et al. [15] solved the deformation flow using a partial differential equation method derived from dynamic optimal transport. Rehman et al. [16] further refined this by introducing a parallelized numerical scheme to facilitate computation. In a different way, Motta et al. [11] and Tian et al. [13] focused on feature-based approaches, extracting graphs from images and applying OT theory to match them. Despite these developments, OT-based registration still faces two fundamental limitations:

1. **Mass preservation assumption.** Most existing methods assume that the masses of intensities or features of two images are equal. When applied to multimodal images, however, this constraint is often violated in multimodal settings due to modality-specific visibility, occlusion, or missing regions.
2. **One-shot correspondence.** Current methods rely on a one-step transport plan to estimate the transformation model. Although some include a mismatch removal step [11], they can still be ineffective if the initial correspondence contains a high proportion of inconsistencies.

Our work. To address these issues, we propose an iterative registration framework based on regularized unbalanced OT, which relaxes the mass constraint and allows progressive refinement of alignment. This strategy significantly improves robustness in the presence of modality-induced inconsistencies, partial overlap, and large deformations.

3. Iterative optimal transport

This section presents our Iterative Optimal Transport (IOT) method. Unlike existing OT-based methods that rely on a one-shot matching step with strict mass preservation assumptions, IOT jointly estimates a transformation and soft correspondences under a relaxed OT framework. First, we introduce our regularized unbalanced optimal transport criterion and transformation models, followed by a discussion of the theoretical properties of the proposed problem. Then, we develop an alternating minimization strategy to iteratively solve the problem and discuss its convergence and complexity.

3.1. Problem formulation

Suppose $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^2$ is a set of points generated from discretizing the edge maps of a moving image, and $\{\boldsymbol{y}_j\}_{j=1}^m \subset \mathbb{R}^2$ is the counterpart from a fixed image. Recall that $\boldsymbol{a}$ and $\boldsymbol{b}$ are their associated histograms, respectively. Without prior knowledge, they are typically chosen as discrete uniform distributions, i.e., $\boldsymbol{a} = n^{-1}\mathbf{1}_n$ and $\boldsymbol{b} = m^{-1}\mathbf{1}_m$. The goal is to align the moving points $\{\boldsymbol{x}_i\}$ with the fixed ones $\{\boldsymbol{y}_j\}$ using a transformation function $f \in \mathcal{H} : \mathbb{R}^2 \to \mathbb{R}^2$.

A natural approach to alignment is to minimize the Wasserstein distance between the distributions of $\{f(\boldsymbol{x}_i)\}$ and $\{\boldsymbol{y}_j\}$, inspired by classical optimal transport theory. However, the classical OT formulation assumes exact mass preservation, which often fails to hold in multimodal settings where structural visibility is often inconsistent. For example, in a retina image pair acquired from different angiographies (Fig. 1), small vessels may appear prominently in one modality but be absent in the other.

In addition, most existing OT-based registration methods adopt a one-shot matching strategy: they compute a single transport plan and use it to estimate the transformation. However, this design lacks a feedback mechanism to iteratively refine the results. Therefore, it can be fragile in scenarios that involve a large initial misalignment or modality-induced distortions. In such cases, the resulting correspondences may be unreliable, leading to poor transformation estimates. As shown in Fig. 1, the transformation $f^{(0)}$ obtained from the initial plan can differ largely from the final optimal solution $f^{(k)}$.

To address such limitations, we adopt the unbalanced OT (UOT) framework [23], which relaxes the strict marginal constraints using KL divergence penalties. This formulation enables soft and partial matching, making it well-suited for multimodal registration. Under this criterion, we jointly estimate the transformation function f and the transport plan $\mathbf{T}$ through an iterative optimization process.

Formally, we minimize the following objective:

$$\min_{f\in\mathcal{H}} \mathrm{UOT}_\lambda(f) + \varepsilon R(f), \tag{2}$$

where the first term,

$$\mathrm{UOT}_\lambda(f) := \min_{\mathbf{T}\in\mathbb{R}_+^{n\times m}} \sum_{i,j} c(f(\boldsymbol{x}_i), \boldsymbol{y}_j) T_{ij} + \lambda\,\mathrm{KL}(\mathbf{T}\mathbf{1}_m \| \boldsymbol{a}) + \lambda\,\mathrm{KL}(\mathbf{T}^\top \mathbf{1}_n \| \boldsymbol{b}), \tag{3}$$

characterizes the similarity between the transformed and fixed points and aims to enforce their closeness; the second term $R(f)$ is a regularizer on f, weighted by the parameter $\varepsilon \geq 0$. Following previous work [24,25], we choose $R(f) = \sum_i \|f(\boldsymbol{x}_i) - \boldsymbol{x}_i\|^2$, ensuring that the moving points do not deviate excessively from their initial positions. The transformation space $\mathcal{H}$ will be elaborated on later.

In the UOT distance defined by (3), we employ the squared Euclidean distance as the ground cost, i.e., $c(f(\boldsymbol{x}_i), \boldsymbol{y}_j) = \|f(\boldsymbol{x}_i) - \boldsymbol{y}_j\|^2$. The cost function can naturally integrate modality-robust features; see Section S1 in the Supplementary Material (SM) for details. The Kullback-Leibler divergence is defined by $\mathrm{KL}(\boldsymbol{p}\|\boldsymbol{q}) = \sum_i p_i \log(p_i/q_i) - p_i + q_i$ with the convention that $0 \log 0 = 0$. The unbalanced relaxation parameter $\lambda > 0$ balances the tolerance for unmatched regions: larger values enforce stricter matching, while smaller values allow more non-overlap between features. As $\lambda \to +\infty$, the formulation degenerates to the classical OT. Compared to the Wasserstein distance, the UOT formulation restricts long-range transportation and allows unassigned mass, thus improving robustness to outliers and partial overlaps resulting from mismatched or missing features. We determine λ adaptively using a robust intensity-based similarity measure; see Section 4.1 for details.

Transformation models. The transformation function f can be either linear or nonlinear. We model f using a global polynomial transformation of degree $q \in \mathbb{Z}_+$:

$$\mathcal{H} = \left\{ f \mid f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x})),\ f_k(\boldsymbol{x}) = \sum_{j=0}^{q} \sum_{i=0}^{j} \theta_{ji}^{(k)} x_1^{j-i} x_2^i \right\}, \quad k = 1, 2. \tag{4}$$

Such a functional space is widely recognized in registration problems for its balance between flexibility and generalization capability [1,6]. Particularly, when the order $q = 1$, the model simplifies to an affine transformation, allowing shifting, scaling, shearing, and rotation to be modeled. For orders $q \geq 2$, it encompasses nonlinear parts to model

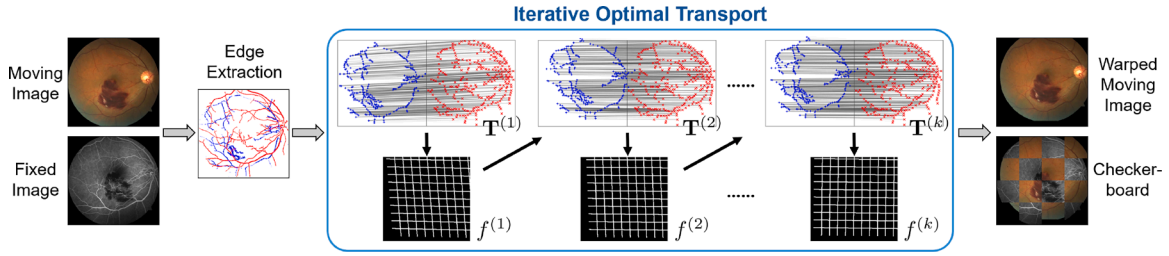


Fig. 1. Overview of the proposed Iterative Optimal Transport method. Given a moving image and a fixed image, structural edges are first extracted from both modalities. At each iteration, a soft correspondence matrix $\mathbf{T}^{(k)}$ between warped moving and fixed points is computed, and a transformation $f^{(k)}$ to warp the moving image is estimated. This process iteratively refines $\mathbf{T}$ and f until the algorithm converges. The final output includes the warped moving image and a checkerboard visualization to assess registration quality.

more complicated deformations. The total number of coefficients in f is $(q+1)(q+2)$, which is independent of the size of the point set, rendering it more computationally efficient compared to those non-rigid methods based on local control points [21]. It's worth noting that other functional spaces, like thin-plate splines (TPS) and reproducing kernel Hilbert space (RKHS), could also be directly integrated into the formulation (2), expanding the versatility of our approach.

Differences from existing UOT-based registration methods. Our work differs from existing UOT-based registration in scope, formulation, and theory. Prior works [12,14] focused on shape or point cloud matching and did not consider the challenges of multimodal image registration. Methodologically, unlike Feydy et al. [12], where UOT serves only as a differentiable fidelity term or Shen et al. [14], where UOT provides one-shot correspondences, we jointly optimize both the transport plan $\mathbf{T}$ and transformation f in a biconvex objective, enabling correspondences to be refined iteratively. Theoretically, we establish biconvexity and convergence guarantees (see Section 3.2), which were not analyzed in the cited works.

3.2. Main algorithm

We use the block coordinate descent (BCD) method [26] to solve the problem (2). The idea is to minimize f and $\mathbf{T}$ alternately for each coordinate, while keeping the other fixed. By exploiting the separability of the feasible region, this method breaks down the original problem into more tractable subproblems. Theorem 1 shows the biconvexity of this problem, with a detailed proof provided in Section S2 of SM.

Theorem 1. *Given* $R(f) = \sum_i \|f(\boldsymbol{x}_i) - \boldsymbol{x}_i\|^2$ *and* $c(f(\boldsymbol{x}_i), \boldsymbol{y}_j) = \|f(\boldsymbol{x}_i) - \boldsymbol{y}_j\|^2$, *the problem* (2) *is biconvex with respect to* f *and* $\mathbf{T}$.

Due to the convexity of the subproblems resulting from BCD, efficient convex minimization strategies can be used to solve these subproblems. The overall alternating minimization procedure is summarized in Algorithm 1 and is illustrated in Fig. 1.

Convergence guarantee. Considering that the objective function of each subproblem in (5) and (6) is continuously differentiable and strictly convex along the corresponding coordinate, according to established results concerning the BCD method [26], the sequence $\{(\boldsymbol{f}^{(k)}, \mathbf{T}^{(k)})\}$ produced by Algorithm 1 is guaranteed to converge to a stationary point.

Computational complexity. In Algorithm 1, we solve the UOT problem in (5) using the maximization-minimization approach proposed by [27], requiring a computational cost of $O(nm)$. For the quadratic optimization problem in (6), we utilize a celebrated quasi-Newton method called Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [28], also demanding a complexity of $O(nm)$. Consequently, the overall time and space complexities of Algorithm 1 are both $O(nm)$, making it well-suited for large-scale image problems.

Extension to 3D images. The proposed IOT method is dimension-agnostic and can be naturally extended to 3D. The main modification is to extract salient 3D structural point sets. This can be done using 3D edge detectors, which are already available in standard libraries. Then, we apply the same BCD framework by alternatively updating the unbalanced OT plan $\mathbf{T}$ and the transformation f in $\mathbb{R}^3$, without modifying the core algorithm.

Algorithm 1 Iterative optimal transport.

1: **Input:** moving points $\{\boldsymbol{x}_i\}_{i=1}^n$, fixed points $\{\boldsymbol{y}_i\}_{i=1}^m$, polynomial order q, regularization parameters ε, λ
2: Initialize $f^{(0)}(\boldsymbol{x}_i) = \boldsymbol{x}_i$, $k = 0$
3: **repeat**
4: Set $k = k + 1$.
5: Update $\mathbf{T}^{(k)}$ by solving the problem

$$\min_{\mathbf{T}\in\mathbb{R}_+^{n\times m}} \sum_{i,j} \|f^{(k-1)}(\boldsymbol{x}_i) - \boldsymbol{y}_j\|^2 T_{ij} + \lambda \,\mathrm{KL}(\mathbf{T}\mathbf{1}_m \| n^{-1}\mathbf{1}_n) + \lambda \,\mathrm{KL}(\mathbf{T}^\top \mathbf{1}_n \| m^{-1}\mathbf{1}_m) \tag{5}$$

with fixed $f^{(k-1)}$ using a maximization-minimization approach [27].
6: Update $f^{(k)}$ by solving the problem

$$\min_{f\in\mathcal{H}} \sum_{i,j} \|f(\boldsymbol{x}_i) - \boldsymbol{y}_j\|^2 T_{ij}^{(k)} + \varepsilon \sum_i \|f(\boldsymbol{x}_i) - \boldsymbol{x}_i\|^2 \tag{6}$$

with fixed $\mathbf{T}^{(k)}$ using a quasi-Newton method [28].
7: **until** Convergence
8: **Output:** $f^{(k)}$, $\mathbf{T}^{(k)}$

4. Experiments

To assess the effectiveness of the proposed Iterative Optimal Transport (IOT) method, we conduct extensive experiments on the registration of diverse multimodal datasets, which are crucial for medical research. We compare IOT with diverse state-of-the-art approaches, considering both qualitative and quantitative evaluations.

4.1. Experimental setup

Datasets. We consider three multimodal image datasets: (i) brain MRI T1- and T2-weighted images (*MRI*) originated from the BrainWeb database[1]; (ii) brain CT and SPECT images (*CT*) originated from the Atlas database[2]; and (iii) retina images from different angiographies (*Retina*) collected by Wang et al. [29]. Each dataset respectively contains 10, 10, and 20 image pairs, each associated with 20 pairs of landmark sets labeled by Jiang et al. [5], which serve as the ground truth for quantitative evaluation. The resolution of the images ranges from 181×217 to 640×640 pixels. The sample data are shown in Fig. 2.

The challenges posed by these datasets progressively increase. For *MRI*, distortions between modalities primarily manifest in colors. On

[1] https://brainweb.bic.mni.mcgill.ca/brainweb/
[2] http://www.med.harvard.edu/aanlib/home.html

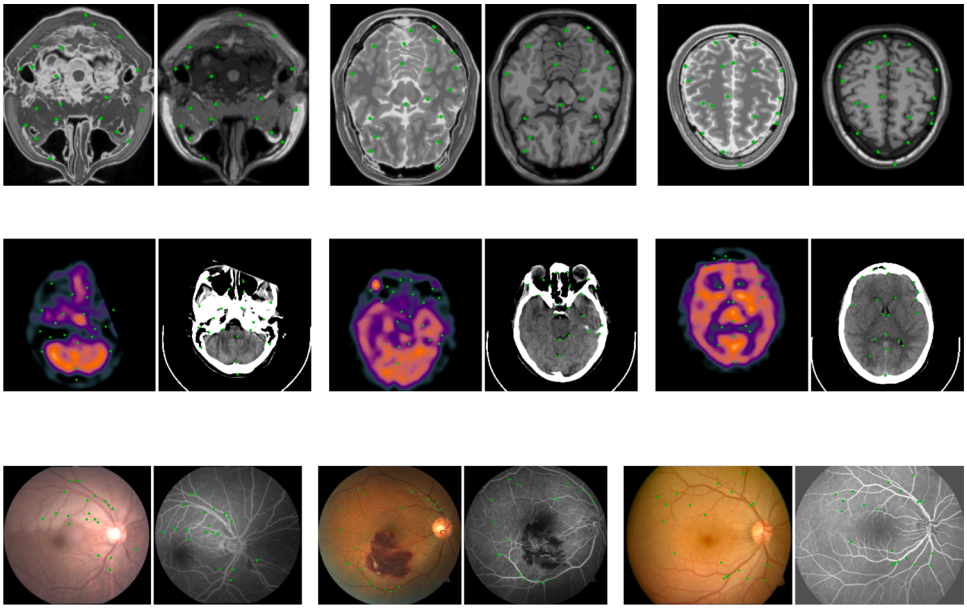


Fig. 2. Sample image pairs in the *MRI* (top row), *CT* (middle row), and *Retina* (bottom row) datasets, where the green dots are corresponding landmarks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the other hand, *CT* exhibits significant differences in texture, resolution, and geometric deformation. *Retina* not only encounters modal differences and deformations, but also has severe partial overlaps. Such diverse scenarios can comprehensively evaluate the validity and robustness of the proposed IOT method.

In order to encompass a variety of geometric deformations, besides aligning real image pairs (referred to as **C1**), we manually build deformations by warping the fixed image using random affine (**C2**) or quadratic (**C3**) transformation matrices following previous literature [1,5,6]. The transformation parameters are uniformly generated within the ranges given in Section S3 of SM. Furthermore, to simulate potential technical errors, we include noisy scenarios where the fixed image is corrupted by random salt-and-pepper noise, accounting for 0.1 % of the pixels.

Competing Methods. We compare IOT with nine baseline methods from different categories as follows. All competing methods are carefully tuned to achieve optimal performance.

1. Point cloud registration methods. We consider two modern point data processing libraries, Open3D [30] and Probreg [31]. Specifically, we use the iterative closest point algorithm from Open3D (Open3D-ICP), and the Gaussian mixture model (Probreg-GMM) and Gaussian filter method (Probreg-Filter) from Probreg, covering representative deterministic and probabilistic approaches.
2. Intensity-based multimodal image registration methods. We include the symmetric normalization method from the ANTs toolbox (ANTs-SyN) [32], which is widely used for medical image registration. We also consider the correlation ratio (CR) via Parzen windowing [33], using publicly available code adapted for 2D image inputs.
3. Feature-based multimodal image registration methods. These include regularized Gaussian fields (RGF) [21], dense adaptive self-correlation (DASC) [34], radiation-variation insensitive feature transform (RIFT) [4], and deformable registration with DINOv2 encoder (DINOreg) [35], covering both hand-crafted and deep learning-based feature representations. All methods are implemented using source code released by their authors, with DINOreg adapted for 2D image registration.

Evaluation. As the metrics used in [4,5], we employ the root mean square error (RMSE) and the mean error (ME) of the matched landmarks to assess the registration accuracy. More precisely, let $\{\boldsymbol{x}_l^*\}_{l=1}^L$ and $\{\boldsymbol{y}_l^*\}_{l=1}^L$ represent the landmark pixels on the moving and fixed images, respectively. Denote $\hat{f}$ as the estimated transformation function. Then, the RMSE and ME for $\hat{f}$ are defined by

$$\mathrm{RMSE} = \sqrt{\frac{1}{L}\sum_{l=1}^{L} \|\hat{f}(\boldsymbol{x}_l^*) - \boldsymbol{y}_l^*\|^2}, \quad \mathrm{ME} = \frac{1}{L}\sum_{l=1}^{L} \|\hat{f}(\boldsymbol{x}_l^*) - \boldsymbol{y}_l^*\|.$$

4.2. Implementation details

Edge maps are extracted from the *MRI* and *CT* datasets using the Canny edge detector, and from the *Retina* dataset using a specialized retinal vessel segmentation algorithm [36]. Subsequently, we apply a sampling method introduced in [37] to discretize these edges into sets of points with $n = m = 300$. Empirical results indicate that the registration performance is not sensitive to the number of points. For a fair comparison, the same points are used in the proposed IOT method, point cloud matching approaches, and the edge-based RGF method. For facilitating parameter selection, we first standardize the feature points such that both $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{y}_j\}$ have zero mean and unit variance.

Hyperparameter selection. The proposed IOT method is influenced by three primary hyperparameters, i.e., q, ε, and λ. The order of polynomials q determines the complexity and flexibility of the transformation model, the regularization parameter ε controls the displacement of moving points, and the marginal relaxation parameter λ indicates the tolerance for mismatch of extracted feature points. We perform a sensitivity analysis based on the *MRI* dataset under **C1** without introducing noise. In particular, the parameters are varied in the range of $q \in \{1, 2, 3\}$, $\varepsilon \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, and $\lambda \in \{10^{-3}, 5\times10^{-3}, 2\times10^{-2}, 10^{-1}, 5\times10^{-1}\}$. The results of RMSE under these varying parameters are depicted in Fig. 3.

The empirical evidence, as illustrated in Fig. 3, suggests that an affine model with proper hyperparameters can provide a satisfactory approximation for actual geometric deformations. Higher-order polynomials exhibit improved robustness to the choice of ε and λ, while they also increase computational demands. To balance these two aspects, we employ the IOT method with affine ($q = 1$, denoted as IOT_1) and quadratic ($q = 2$, denoted as IOT_2) transformations in the subsequent experiments.

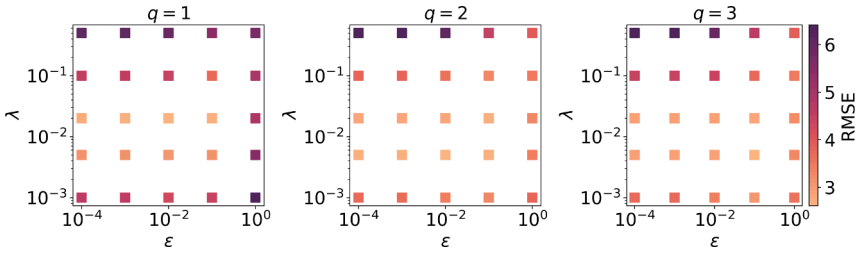


Fig. 3. Effect of the order of polynomials q, the regularization parameter ε, and the marginal relaxation parameter λ on RMSE for the proposed IOT method. The average values on the *MRI* dataset are reported.

Moreover, the IOT method shows superior performance over a wide range of ε. Specifically, as long as ε is not too large, the resulting RMSE is approximately in the same order. In contrast, very large values of ε tend to bias the solution too far away from the UOT paradigm. Consequently, we set the regularization parameter to $\varepsilon = 10^{-2}$.

Compared to ε, we observe that the IOT method is more sensitive to the marginal relaxation parameter λ. Too large λ can force the alignment of unmatched feature points, resulting in high registration errors. This highlights the necessity of unbalanced relaxation in the IOT method. On the other hand, exceedingly small values of λ would overly loosen the marginal constraints, leading to a solution that deviates from the true underlying transport structure. In general, the optimal λ depends on both the degree of non-overlap and feature mismatch in image pairs, which can vary significantly across different images. Therefore, instead of fixing it in advance, we automatically tune this parameter by searching among $\lambda \in \{10^{-3}, 5\times10^{-3}, 2\times10^{-2}, 10^{-1}, 5\times10^{-1}\}$ and choose the one that maximizes the cross cumulative residual entropy [38,39] between the warped moving image and fixed image.

4.3. Performance comparison

To evaluate our proposed IOT method comprehensively, we compare it with other state-of-the-art registration methods through both quantitative and qualitative aspects of registration accuracy, as well as an assessment of computational efficiency.

Quantitative performance. We evaluate the registration accuracy of our IOT method, specifically IOT_1 and IOT_2, by comparing with the baseline methods mentioned above. Since errors across different image pairs may vary dramatically, we use rank instead of the original error for performance comparison [40]. More specifically, for each image pair, all methods are ranked based on their RMSE, with the best-performing method receiving rank 1 and the poorest performer receiving rank 11. Then we average the ranks of a method across all image pairs in a dataset. Such average ranks based on RMSE are presented in Tables 1–3. The results based on ME, which display a similar pattern, are included in Section S4.1 of SM.

Overall, we observe that the proposed IOT method consistently achieves the smallest RMSE ranking. In almost all cases, with and without added noise, IOT_1 and IOT_2 rank in the top three positions, indicating superior registration accuracy and robustness. We also observe that in simple cases with negligibly small initial registration errors (e.g., **C1** in *MRI* dataset; as illustrated in Fig. 2), the competing approaches, including Probreg-GMM (point cloud), CR (intensity-based), and RGF/DINOreg (feature-based) also perform well, with comparable accuracy to our IOT. However, their performance deteriorates notably when noise or deformation complexity increases, as observed from **C1** to **C2** in Table 1 for Probreg-GMM/CR/DINOreg, and from the *MRI* dataset (Table 1) to the *CT* dataset (Table 2) for RGF.

In Tables 1 and 2, the Probreg-GMM method, although underperforming our IOT method, shows competitive performance on the brain image datasets; however, its efficacy diminishes when applied to the more intricate *Retina* dataset, as shown in Table 3. In comparison, our approach consistently yields high registration accuracy across a variety of image types. Another notable competitor is the RGF method, which also achieves a relatively low average rank in a majority of instances. Nevertheless, its performance is unstable, as reflected in its standard deviations, leading to inferior registration results at times. The deep learning-based DINOreg method also achieves moderate success but does not reach the top ranks of IOT.

Additionally, we note that IOT_2 may exhibit inferior performance compared to IOT_1 when modeling non-rigid deformation under noisy conditions (e.g., **C3** with noise in Table 1). This can be attributed to the selected edge maps being corrupted by outliers, which result from

Table 1
Average rank (with standard deviation in parentheses) based on RMSE of each method on the *MRI* dataset in the cases of C1–C3, without or with noise. The smaller the better.

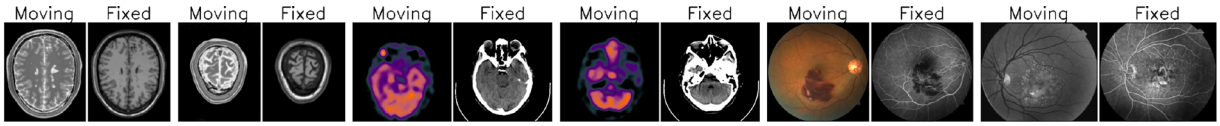
Method	C1		C2		C3	
	w/o noise	w/ noise	w/o noise	w/ noise	w/o noise	w/ noise
Open3D-ICP	5.0 (1.9)	5.0 (2.5)	4.3 (0.8)	4.8 (0.7)	4.9 (0.3)	4.8 (0.4)
Probreg-Filter	11.0 (0.0)	11.0 (0.0)	6.1 (1.4)	5.8 (1.2)	6.9 (2.8)	5.9 (2.8)
Probreg-GMM	***2.3*** (0.6)	***2.3*** (0.9)	4.6 (0.9)	3.5 (1.4)	**3.6** (0.5)	**3.0** (0.9)
ANTs-SyN	7.8 (2.3)	8.3 (0.9)	7.8 (0.4)	8.0 (1.1)	8.4 (1.1)	8.5 (1.1)
CR	4.9 (2.8)	3.0 (2.5)	10.2 (0.9)	9.8 (1.5)	8.8 (2.1)	9.6 (1.9)
RGF	***2.3*** (0.6)	***2.3*** (0.9)	**2.7** (1.2)	**3.3** (2.4)	**3.6** (1.6)	4.2 (2.5)
DASC	10.0 (0.0)	10.0 (0.0)	9.7 (1.2)	9.9 (1.0)	9.7 (1.1)	9.7 (0.8)
RIFT	7.1 (3.1)	8.5 (0.5)	9.5 (1.6)	9.0 (1.7)	8.4 (1.7)	8.1 (1.9)
DINOreg	4.2 (2.9)	4.2 (2.8)	7.3 (0.8)	7.6 (0.9)	8.2 (0.9)	7.8 (1.1)
IOT_1	***2.3*** (0.6)	***2.3*** (0.9)	***1.3*** (0.5)	***1.7*** (0.8)	**2.1** (0.9)	***1.4*** (0.5)
IOT_2	***2.3*** (0.6)	***2.3*** (0.9)	**2.5** (0.9)	**2.6** (0.9)	***1.4*** (0.5)	**3.0** (0.8)

* The top-3 results of each case are in bold, and the best result is in italics.
* The median initial RMSEs for **C1**–**C3** are 0.97, 29.72, and 34.64, respectively.

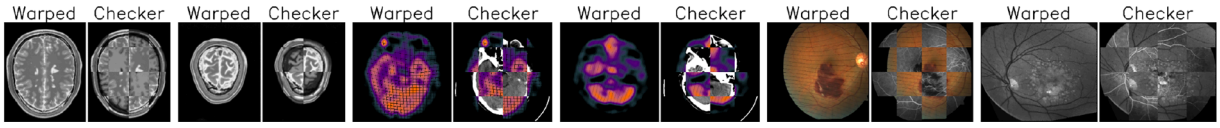
Table 2
Average rank (with standard deviation in parentheses) based on RMSE of each method on the *CT* dataset in the cases of **C1**–**C3**, without or with noise. The smaller the better.

Method	C1		C2		C3	
	w/o noise	w/ noise	w/o noise	w/ noise	w/o noise	w/ noise
Open3D-ICP	**3.7** (1.0)	5.3 (1.5)	4.3 (0.9)	4.7 (0.9)	4.5 (0.7)	5.0 (1.0)
Probreg-Filter	8.3 (1.7)	9.8 (1.8)	6.5 (2.8)	5.5 (2.6)	6.6 (2.7)	5.6 (3.2)
Probreg-GMM	4.6 (1.7)	4.5 (1.6)	**3.4** (1.0)	**2.7** (0.9)	**2.6** (1.2)	***2.3*** (1.3)
ANTs-SyN	6.0 (0.8)	5.3 (1.6)	7.8 (1.0)	7.9 (0.5)	8.2 (1.5)	8.5 (1.4)
CR	10.5 (0.7)	9.8 (1.2)	9.2 (1.8)	10.1 (1.0)	9.0 (1.5)	8.0 (1.9)
RGF	5.4 (2.6)	5.3 (2.7)	4.4 (2.1)	4.7 (1.6)	4.5 (2.3)	4.8 (2.4)
DASC	9.3 (1.1)	8.7 (0.8)	10.6 (0.5)	10.4 (0.5)	10.3 (0.8)	10.5 (0.7)
RIFT	9.1 (1.0)	8.8 (1.2)	8.5 (1.7)	8.5 (1.7)	7.4 (2.7)	7.7 (2.4)
DINOreg	5.6 (1.2)	**4.4** (1.9)	7.0 (1.0)	7.1 (0.5)	7.8 (1.5)	8.1 (1.0)
IOT_1	***1.1*** (0.3)	***1.6*** (1.5)	***1.4*** (0.7)	***2.0*** (1.5)	**2.7** (1.4)	**2.4** (1.4)
IOT_2	**2.4** (0.8)	**2.5** (1.0)	**2.9** (1.4)	**2.4** (1.1)	***2.4*** (1.1)	**3.1** (0.9)

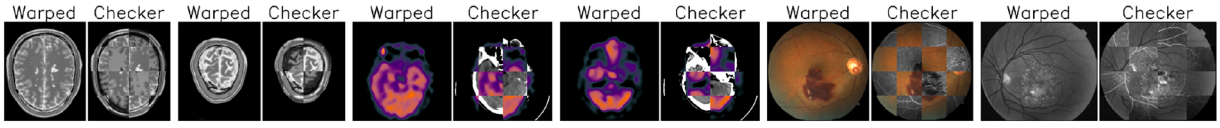
* The top-3 results of each case are in bold, and the best result is in italics.
* The median initial RMSEs for **C1**–**C3** are 14.64, 39.66, and 40.95, respectively.



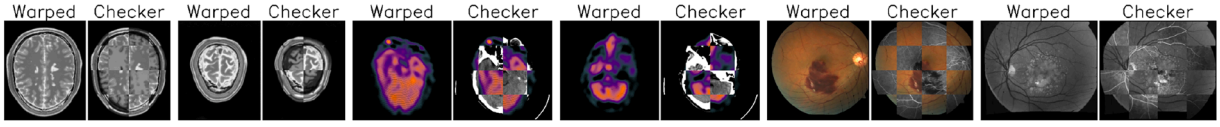
(a) Original image pairs.



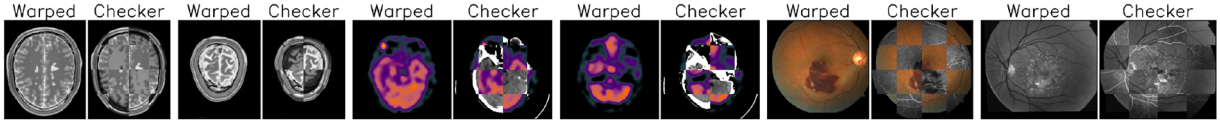
(b) Registration results of the Probreg-GMM method.



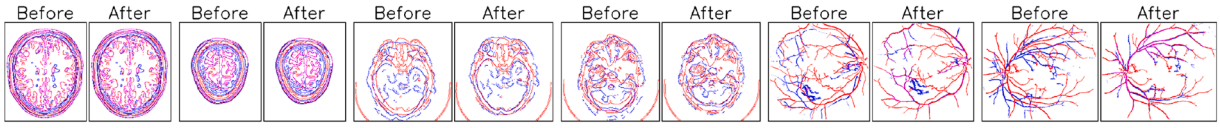
(c) Registration results of the ANTs-SyN method.



(d) Registration results of the RGF method.



(e) Registration results of the IOT method.



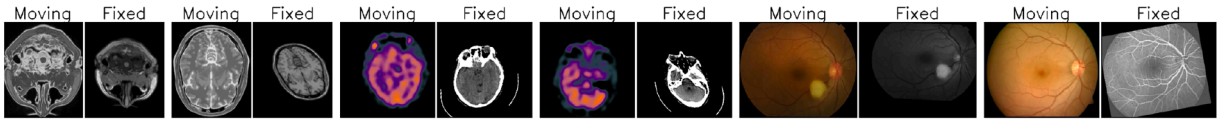
(f) Original and warped edge maps using the IOT method.

Fig. 4. Qualitative comparison of different methods in **C1**. For each pair in Fig. 4(a), the left is the moving image, and the right is the fixed image. For each pair in Figs. 4(b)–(e), the left is the warped moving image yielded by each method, and the right is the checkerboard of warped moving and fixed images. For each pair in Fig. 4(f), the left represents the edge maps before registration, and the right represents the edge maps after registration using the proposed IOT method, where the edges of the moving image are in blue and those of the fixed image are in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

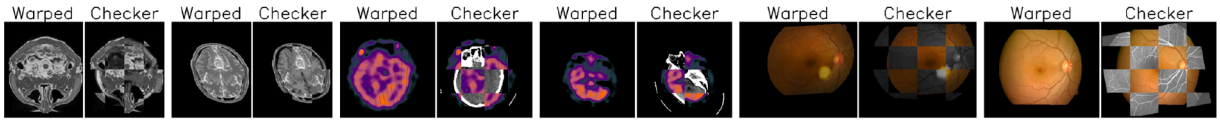
image noise. An affine model, which broadly aligns contours, tends to produce relatively small matching errors. By contrast, a non-rigid model may encourage aligning some outliers to minimize the registration criterion. This misalignment can lead to overfitting on outliers, reducing the registration accuracy. Therefore, we recommend opting for $q = 1$ in the transformation function in the absence of severe non-rigid deformations.

Qualitative performance. We next conduct qualitative comparisons on the three datasets. For clarity, we focus on only the top-performing competitors within each category, i.e., Probreg-GMM, ANTs-SyN, and RGF, alongside our IOT method for qualitative analyses. We randomly select two image pairs from each dataset for illustration. The original images and registration results are presented in Figs. 4–6, corresponding to cases **C1**–**C3**, respectively. Here, the warped moving images are overlaid onto the fixed images using a checkerboard pattern. Additionally, for the IOT method, we also display the warped edge maps to show its effectiveness.

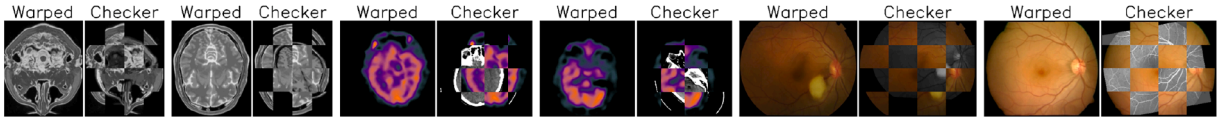
Figs. 4–6 clearly reveal considerable variations in texture, color, and lightness across different modalities. Furthermore, there are significant areas of non-overlap in images (e.g., **Retina** images in Fig. 6) and a large



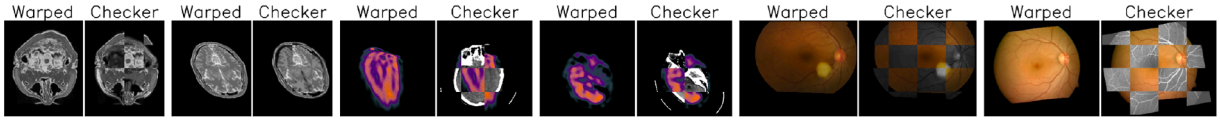
(a) Original image pairs.



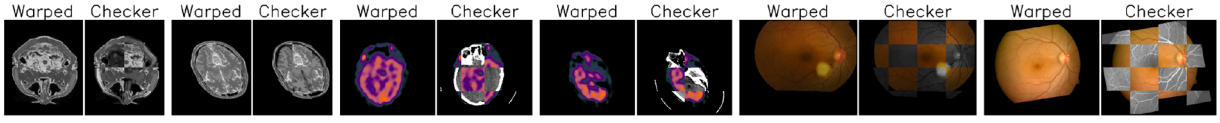
(b) Registration results of the Probreg-GMM method.



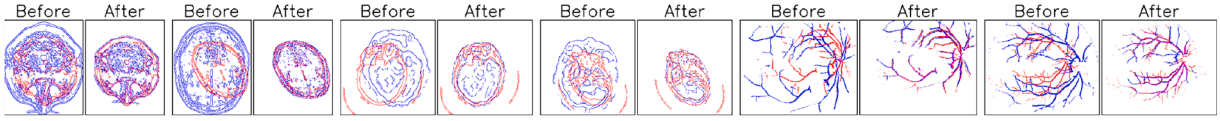
(c) Registration results of the ANTs-SyN method.



(d) Registration results of the RGF method.



(e) Registration results of the IOT method.



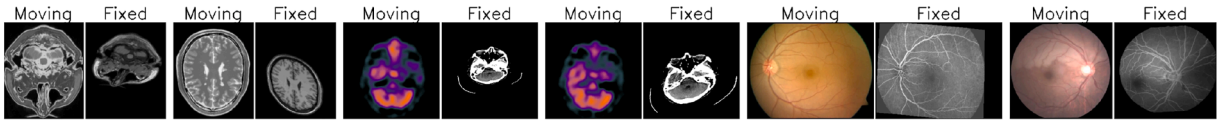
(f) Original and warped edge maps using the IOT method.

Fig. 5. Qualitative comparison of different methods in **C2**. For each pair in Fig. 5(a), the left is the moving image, and the right is the fixed image. For each pair in Figs. 5(b)–(e), the left is the warped moving image yielded by each method, and the right is the checkerboard of warped moving and fixed images. For each pair in Fig. 5(f), the left represents the edge maps before registration, and the right represents the edge maps after registration using the proposed IOT method, where the edges of the moving image are in blue and those of the fixed image are in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

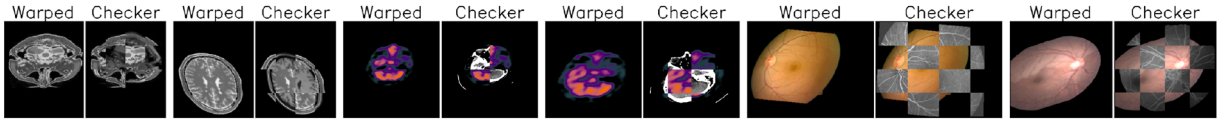
proportion of mismatched feature points (e.g., **CT** images in Fig. 5). In such challenging scenarios, our IOT method demonstrates superior accuracy compared to other leading approaches, as evident in both checkerboards and overlap edges. Taking Fig. 5 as an example, the ME of IOT on each image pair equals 1.71, 2.54, 4.29, 4.14, 2.78, and 2.94, respectively. These results further confirm that our method can achieve precise registration for multimodal images, even in the face of substantial modal differences, severe geometric deformations, and extensive non-overlapping regions.

The observed robustness of IOT to noisy or mismatched edge sets arises from three key designs. First, we estimate correspondences within the algorithm via unbalanced OT, which allows partial matching, so modality-specific or missing structures are naturally left unmatched instead of being forced to align. Second, we re-estimate the transport plan and the transformation iteratively. Early iterations coarsely align geometry, and subsequent iterations progressively refine correspondences, thereby reducing sensitivity to local edge misalignment. Third, we adjust the marginal relaxation content-adaptive, so the amount of unmatched mass is tuned to image statistics, further improving robustness when the shareable structure is limited.

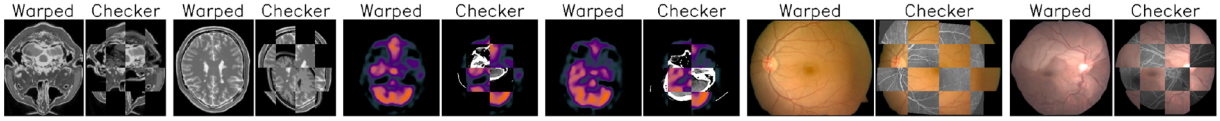
Although IOT is motivated by multimodal challenges, the method itself is not restricted to cross-modality settings. To demon-



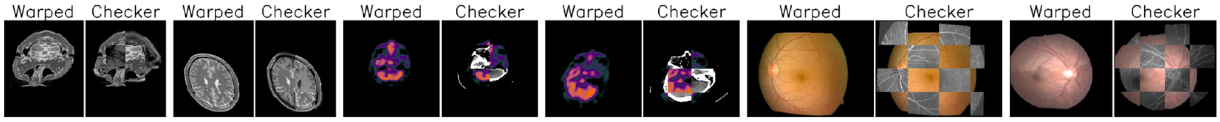
(a) Original image pairs.



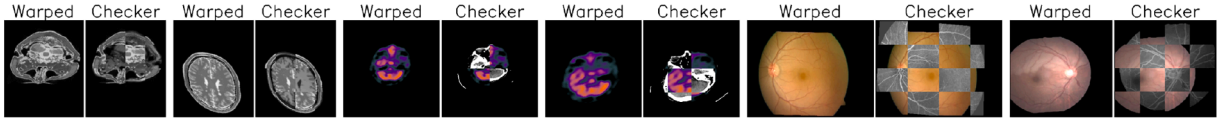
(b) Registration results of the Probreg-GMM method.



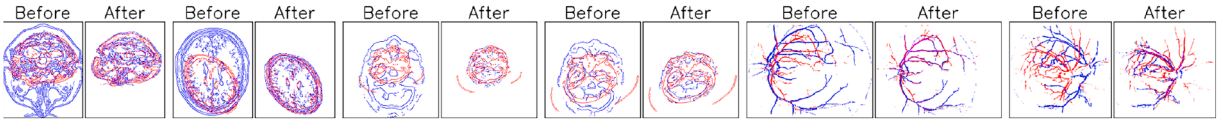
(c) Registration results of the ANTs-SyN method.



(d) Registration results of the RGF method.



(e) Registration results of the IOT method.



(f) Original and warped edge maps using the IOT method.

Fig. 6. Qualitative comparison of different methods in **C3**. For each pair in Fig. 6(a), the left is the moving image, and the right is the fixed image. For each pair in Figs. 6(b)–(e), the left is the warped moving image yielded by each method, and the right is the checkerboard of warped moving and fixed images. For each pair in Fig. 6(f), the left represents the edge maps before registration, and the right represents the edge maps after registration using the proposed IOT method, where the edges of the moving image are in blue and those of the fixed image are in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

strate broader applicability, we additionally evaluated IOT on unimodal benchmarks; see Section S4.2 of SM for details and results.

Running time. Table 4 reports the average computational time of each registration method on the three datasets. IOT_1 performs at a mid-tier level in terms of computational efficiency, computing faster than other feature-based registration approaches, while requiring a longer but comparable runtime when compared to most point cloud matching and intensity-based image registration algorithms. IOT_2 is slightly less efficient than IOT_1 due to the additional complexity in estimating non-rigid transformations.

The computational time of the IOT method may be reduced. We find that its runtime is mainly dominated by solving the UOT subproblem. To improve the efficiency, we can leverage recently advanced fast UOT solvers that approximate solutions within a linear time. Additionally, rewriting the code in C++ could significantly reduce the runtime, potentially by an order of magnitude. Furthermore, exploring GPU parallel processing also presents a promising route for further accelerating the computation.

Beyond accuracy and efficiency, we also evaluated the physical plausibility of the estimated transformations using Jacobian-based regularity metrics; see Section S4.3 of SM for detailed results.

Table 3
Average rank (with standard deviation in parentheses) based on RMSE of each method on the *Retina* dataset in the cases of **C1**–**C3**, without or with noise. The smaller the better.

Method	C1		C2		C3	
	w/o noise	w/ noise	w/o noise	w/ noise	w/o noise	w/ noise
Open3D-ICP	5.9 (2.1)	6.0 (1.9)	**4.2** (1.7)	4.8 (1.8)	**3.1** (2.5)	**3.7** (2.3)
Probreg-Filter	8.3 (1.8)	7.9 (2.0)	7.8 (2.0)	8.1 (1.9)	7.0 (1.9)	5.8 (2.7)
Probreg-GMM	7.5 (2.3)	6.1 (2.9)	6.7 (2.4)	6.3 (3.2)	7.0 (2.1)	6.3 (2.6)
ANTs-SyN	5.1 (2.3)	5.4 (2.5)	4.8 (2.6)	5.1 (2.7)	5.7 (2.3)	5.9 (2.2)
CR	8.9 (2.3)	8.9 (1.8)	10.1 (1.4)	8.7 (3.2)	9.1 (2.9)	9.3 (1.7)
RGF	***2.8*** (2.7)	**3.6** (3.1)	5.4 (3.7)	**4.0** (3.3)	4.9 (3.5)	5.4 (4.1)
DASC	8.6 (2.5)	8.7 (2.3)	8.7 (1.9)	9.0 (1.8)	9.5 (1.6)	8.8 (2.5)
RIFT	7.5 (3.1)	7.9 (3.3)	6.3 (3.4)	6.7 (3.5)	6.6 (3.3)	7.0 (3.2)
DINOreg	4.9 (2.4)	4.9 (2.5)	4.6 (2.8)	5.3 (2.8)	5.8 (1.9)	6.1 (2.7)
IOT_1	**3.0** (1.6)	***3.1*** (2.0)	***3.1*** (1.6)	***3.4*** (2.0)	***2.6*** (1.7)	***3.3*** (2.0)
IOT_2	**3.5** (1.5)	**3.5** (1.6)	**4.3** (1.8)	**4.6** (1.0)	**4.7** (1.5)	**4.4** (1.8)

* The top-3 results of each case are in bold, and the best result is in italics.
* The median initial RMSEs for **C1**–**C3** are 20.66, 36.82, and 36.84, respectively.

Table 4
Average computational time (in seconds) per image pair of each method on three datasets.

Method	*MRI*	*CT*	*Retina*
Open3D-ICP	2.81	2.91	4.03
Probreg-Filter	2.82	2.93	4.04
Probreg-GMM	3.69	4.14	6.02
ANTs-SyN	1.48	2.67	4.84
CR	9.85	12.76	17.70
RGF	8.54	12.71	15.01
DASC	17.33	25.88	33.35
RIFT	13.42	16.04	19.78
DINOreg	50.14	96.12	161.63
IOT_1	4.85	5.72	7.43
IOT_2	6.48	8.14	9.67

5. Conclusion

In this study, we have introduced a novel and robust Iterative Optimal Transport (IOT) approach for multimodal image registration. The key idea is to minimize a regularized UOT criterion, which we address through alternatively solving for the transformation function and the correspondence relationship, backed by theoretical convergence guarantees. Through extensive testing, our IOT method has shown its superiority in handling a diverse range of modal discrepancies, geometric distortions, and noise levels. While the experiments mainly focus on medical images, our method also yields promising performance in a broader spectrum of scenarios, like visible and infrared image registration, further demonstrating its impressive adaptability and universality to various registration challenges among different modalities.

Although our design reduces dependence on the quality of extracted edge sets, we acknowledge a natural limitation: performance may degrade when the two images lack sufficiently salient shared structures. For example, abdominal MRI-CT image pairs frequently exhibit ambiguous organ boundaries, large modality-specific regions, and local non-rigid deformations. These factors make our edge-driven global registration insufficient. To address this, a more reliable organ extractor should replace general edge detection, and then the feature-augmented cost can provide more meaningful semantic guidance. Furthermore, IOT can be implemented in a coarse-to-fine manner. This approach begins with a global model and progressively incorporates more expressive deformation families, such as TPS or RKHS, at finer scales to capture complex local non-rigid motions. We leave these directions to future work.

Finally, while the computational intensity remains a challenge for IOT, we have discovered potential directions for acceleration, which will be a focus of our future work. Moreover, we plan to utilize this powerful registration tool for downstream analyses such as visual tracking and multimodal image fusion, paving the way for both exploratory research and practical applications.

CRediT authorship contribution statement

Mengyu Li: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis; **Cheng Meng:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization; **Xiaodan Fan:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Data curation, Conceptualization.

Data availability

I have shared the link to the data/code in our manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.patcog.2025.112736.

References

[1] C. Min, Y. Gu, Y. Li, F. Yang, Non-rigid infrared and visible image registration by enhanced affine transformation, Pattern Recognit. 106 (2020) 107377.
[2] X. Zhang, P. Ye, H. Leung, K. Gong, G. Xiao, Object fusion tracking based on visible and infrared images: a comprehensive review, Inf. Fusion 63 (2020) 166–187.
[3] J. Ma, X. Jiang, A. Fan, J. Jiang, J. Yan, Image matching from handcrafted to deep features: a survey, Int. J. Comput. Vis. 129 (2021) 23–79.
[4] J. Li, Q. Hu, M. Ai, RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform, IEEE Trans. Image Process. 29 (2020) 3296–3310.
[5] X. Jiang, J. Ma, G. Xiao, Z. Shao, X. Guo, A review of multimodal image matching: methods and applications, Inf. Fusion 73 (2021) 22–71.
[6] M. Hasan, M.R. Pickering, X. Jia, Robust automatic registration of multimodal satellite images using CCRE with partial volume interpolation, IEEE Trans. Geosci. Remote Sens. 50 (10) (2012) 4050–4061.
[7] Z. Soleimanitaleb, M.A. Keyvanrad, A. Jafari, Object tracking methods: a review, in: 2019 9th International Conference on Computer and Knowledge Engineering, IEEE, 2019, pp. 282–288.
[8] H. Xu, J. Yuan, J. Ma, MURF: mutually reinforcing multi-modal image registration and fusion, IEEE Trans. Pattern Anal. Mach. Intell. 45 (10) (2023) 12148–12166.
[9] C. Villani, Topics in Optimal Transportation, 58, American Mathematical Society, 2021.
[10] J. Zhang, W. Zhong, P. Ma, A review on modern computational optimal transport methods with applications in biomedical research, Mod. Stat. Methods Health Res. (2021) 279–300.
[11] D. Motta, W. Casaca, A. Paiva, Vessel optimal transport for automated alignment of retinal fundus images, IEEE Trans. Image Process. 28 (12) (2019) 6154–6168.
[12] J. Feydy, B. Charlier, F.-X. Vialard, G. Peyré, Optimal transport for diffeomorphic registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 291–299.
[13] X. Tian, N. Anantrasirichai, L. Nicholson, A. Achim, Optimal transport-based graph matching for 3D retinal OCT image registration, in: 2022 IEEE International Conference on Image Processing, IEEE, 2022, pp. 2791–2795.
[14] Z. Shen, J. Feydy, P. Liu, A.H. Curiale, R. San Jose Estepar, R. San Jose Estepar, M. Niethammer, Accurate point cloud registration with robust optimal transport, Adv. Neural Inf. Process. Syst. 34 (2021) 5373–5389.
[15] S. Haker, L. Zhu, A. Tannenbaum, S. Angenent, Optimal mass transport for registration and warping, Int. J. Comput. Vis. 60 (2004) 225–240.
[16] T.u. Rehman, E. Haber, G. Pryor, J. Melonakos, A. Tannenbaum, 3D nonrigid registration via optimal mass transport on the GPU, Med. Image Anal. 13 (6) (2009) 931–940.
[17] B.B. Avants, C.L. Epstein, M. Grossman, J.C. Gee, Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain, Med. Image Anal. 12 (1) (2008) 26–41.
[18] D. Sengupta, P. Gupta, A. Biswas, A survey on mutual information based medical image registration algorithms, Neurocomputing 486 (2022) 174–188.
[19] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.

[20] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, IEEE, 2005, pp. 886–893.
[21] J. Ma, J. Zhao, Y. Ma, J. Tian, Non-rigid visible and infrared face registration via regularized Gaussian fields criterion, Pattern Recognit. 48 (3) (2015) 772–784.
[22] G. Haskins, U. Kruger, P. Yan, Deep learning in medical image registration: a survey, Mach. Vis. Appl. 31 (2020) 1–18.
[23] L. Chizat, G. Peyré, B. Schmitzer, F.-X. Vialard, Scaling algorithms for unbalanced optimal transport problems, Math. Comput. 87 (314) (2018) 2563–2609.
[24] S.G. Kong, J. Heo, F. Boughorbel, Y. Zheng, B.R. Abidi, A. Koschan, M. Yi, M.A. Abidi, Multiscale fusion of visible and thermal IR images for illumination-invariant face recognition, Int. J. Comput. Vis. 71 (2007) 215–233.
[25] M. Perrot, N. Courty, R. Flamary, A. Habrard, Mapping estimation for discrete optimal transport, Adv. Neural Inf. Process. Syst. 29 (2016) 4204–4212.
[26] D.P. Bertsekas, Nonlinear Programming, 3rd Edition, Athena Scientific, 2016.
[27] L. Chapel, R. Flamary, H. Wu, C. Févotte, G. Gasso, Unbalanced optimal transport through non-negative penalized linear regression, Adv. Neural Inf. Process. Syst. 34 (2021) 23270–23282.
[28] J. Nocedal, S.J. Wright, Numerical Optimization, 2rd Edition, Springer, 2006.
[29] G. Wang, Z. Wang, Y. Chen, W. Zhao, Robust point matching method for multimodal retinal image registration, Biomed. Signal Process. Control 19 (2015) 68–76.
[30] Q.-Y. Zhou, J. Park, V. Koltun, Open3D: a modern library for 3D data processing, (2018). arXiv preprint arXiv:1801.09847
[31] K. Tanaka, P. Schmitz, M. Ciganovic, P. Kumar, Probreg: probablistic point cloud registration library, 2020, (https://probreg.readthedocs.io/en/latest/).
[32] N.J. Tustison, P.A. Cook, A.J. Holbrook, H.J. Johnson, J. Muschelli, G.A. Devenyi, J.T. Duda, S.R. Das, N.C. Cullen, D.L. Gillen, et al., The ANTsX ecosystem for quantitative biological and medical imaging, Sci. Rep. 11 (1) (2021) 9068.
[33] J. Chen, Y. Liu, S. Wei, A. Carass, Y. Du, Unsupervised learning of multi-modal affine registration for PET/CT, in: 2024 IEEE Nuclear Science Symposium (NSS), Medical Imaging Conference (MIC) and Room Temperature Semiconductor Detector Conference (RTSD), IEEE, 2024, pp. 1–2.
[34] S. Kim, D. Min, B. Ham, M.N. Do, K. Sohn, DASC: robust dense descriptor for multi-modal and multi-spectral correspondence estimation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (9) (2017) 1712–1729.
[35] X. Song, X. Xu, J. Zhang, D.M. Reyes, P. Yan, Dino-Reg: efficient multimodal image registration with distilled features, IEEE Trans. Med. Imaging 44 (9) (2025) 3809–3819.
[36] T. Coye, A novel retinal blood vessel segmentation algorithm for fundus images, Matlab Cent. File Exch. (2015). https://ww2.mathworks.cn/matlabcentral/fileexchange/50839-novel-retinal-vessel-segmentation-algorithm-fundus-images
[37] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 509–522.
[38] F. Wang, B.C. Vemuri, M. Rao, Y. Chen, A new & robust information theoretic measure and its application to image alignment, in: Biennial International Conference on Information Processing in Medical Imaging, Springer, 2003, pp. 388–400.
[39] M. Rao, Y. Chen, B.C. Vemuri, F. Wang, Cumulative residual entropy: a new measure of information, IEEE Trans. Inf. Theory 50 (6) (2004) 1220–1228.
[40] Z. Liu, D. Wu, W. Zhai, L. Ma, SONAR Enables cell type deconvolution with spatially weighted Poisson-Gamma model for spatial transcriptomics, Nat. Commun. 14 (1) (2023) 4727.