FULLY DIFFERENTIABLE CORRELATION-DRIVEN 2D/3D REGISTRATION FOR X-RAY TO CT IMAGE FUSION

Minheng Chen[1] *Zhirun Zhang*[1] *Shuheng Gu*[1] *Zhangyang Ge*[1] *Youyong Kong*[123]

[1] School of Computer Science and Engineering, Southeast University, China
[2] Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, Southeast University, China
[3] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, Nanjing, China

ISBI 2024 ATHENS, GREECE

INTRODUCTION

• Background

Rigid 2D/3D registration is crucial for merging preoperative CT with live X-ray images in minimally invasive surgeries. Despite the accuracy of traditional optimization methods, they have limitations due to sensitivity to initialization and image mismatch. Recent deep learning advancements offer a fully differentiable approach to enhance registration performance and capture range. Inspired by previous optimization-based Methods using gradient correlation as a similarity metric for registration, low-frequency information tends to introduce significant interference when the estimated pose is close to ground truth, while it can increase capture range with global information when the estimated pose far away from ground truth.

• Main Challenges

1) Lack of Controllability and Interpretability
2) Limited Capture Range

• Contribution

1) **Dual-Branch Encoder Innovation**: We have developed a dual-branch CNN-Transformer encoder that enables the extraction and separation of both local and global features from the imagery data.
2) **Correlation-Driven Loss Function**: We have introduced a novel loss function that is driven by correlation, which is specifically designed for the decomposition of low-frequency and high-frequency features.
3) **Training Strategy for Convex-Shaped Similarity**: We adopt a training strategy that aims to learn an approximation of a convex-shaped similarity function.
4) **Superior Framework Demonstration**: We have provided evidence that demonstrates the superior performance of our correlation-driven framework, particularly on synthetic data, showcasing its robustness and potential in practical applications.

• Problem Definition

$$\mathcal{F}(\theta) = arg\min_{\theta} \mathcal{S}(I_x, I_m)$$
$$= arg\min_{\theta} \mathcal{S}(I_x, P(\theta; V))$$

where S represents a similarity metric between the intraoperative fluoroscopic image I_x and the DRR I_m. $P(\theta; V)$ denotes the generation of DRR from volumetric 3D scene V by using a 6 DoF pose θ and projection operator P.

METHODOLOGY

• Network Architecture

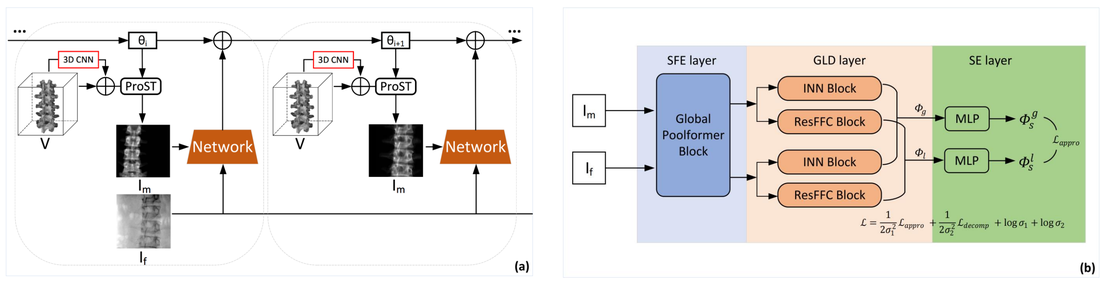

Figure. 1(a) shows the architecture of the proposed framework. Given an input volume $V \in R^{H\times W\times D}$, a fixed 2D image $I_x \in R^{H\times W}$ and an initial pose $\theta_{ini} \in SE(3)$, where H, W and D denote the height, width, and depth, respectively. We employ a 3D CNNto learn the residual from V. The projected moving image $I_m$is generated by using the ProST projection module. And then I_m and I_x will pass through the proposed dual-branch CNN-Transformer encoder. The structure of the encoder is illustrated in Fig.1(b), which consists of three components: the shallow share feature encoder(SFE), the global-local feature decomposition(GLD) layer and the similarity evaluation (SE) layer.

- Shallow share feature encoder: The objective of SFE is to extract shallow features from I_m and I_x individually.
- Global-local feature decomposition layer: The GLD layer aims to extract and decouple global and local features from the shared features Φ_s^x and Φ_s^m .We adopt the INN block with affine coupling layers. Let the global feature extraction and local feature extraction be represented by G(·) and L(·) respectively: $\Phi_g = (G(\Phi_s^x) - G(\Phi_s^m))^2, \Phi_l = (L(\Phi_s^x) - L(\Phi_s^m))^2$
- Similarity evaluation layer. The function of the SE layer is to estimate image similarity on local/global features Φ_l and Φ_g, respectively.

• Loss Function

Specifically, at training stage, the gradient approximation loss $\mathcal{L}$ is(where $\mathcal{L}_{net}$ is the loss of the network):

$$\mathcal{L}_{appro} = \mathcal{L}_{geo}(\frac{\partial \mathcal{L}_{net}}{\partial \theta}, \frac{\partial \mathcal{L}_{geo}(\theta, \theta_t)}{\partial \theta})$$

$$\mathcal{L}_{net} = \langle \sigma(\Phi_g^s), \sigma(\Phi_l^s) \rangle$$

we propose a correlation-based decomposition loss Ldecomp which uses the normalized cross correlation operator NCC(·, ·) to decouple the local/global information.

$$\mathcal{L}_{decomp} = \frac{\mathrm{NCC}(G(\Phi_s^x), G(\Phi_s^m))}{\mathrm{NCC}(L(\Phi_s^x), L(\Phi_s^m)) + \epsilon}$$

However, achieving a balance in the weights of multiple losses is still a challenging task, even when employing the time-consuming grid search methods to tune hyper-parameters. This makes the gradient flow transmission process difficult to control. As a result, we opt for a loss function that incorporates uncertain weights during training:

$$\mathcal{L} = \frac{1}{2\sigma_1^2}\mathcal{L}_{appro} + \frac{1}{2\sigma_2^2}\mathcal{L}_{decomp} + \log\sigma_1 + \log\sigma_2$$

where σ1, σ2 are learnable variables and ϵ is hyperparameter

• Inference Phase

Since our network is fully differentiable, the registration can be viewed as a gradient-based iterative optimization process. The output of the well-trained network ϕ with fixed parameters can be considered as a similarity objective function. And the i-th stage of this iterative alignment can be shown as:

$$\theta_i = \theta_{i-1} - \alpha \frac{\partial \phi(\theta_{i-1})}{\partial I_m^{i-1}} \frac{\partial I_m^{i-1}}{\partial P(\theta_{i-1}; V)} \frac{\partial P(\theta_{i-1}; V)}{\partial \theta_{i-1}}$$

EXPERIMENT

• Dataset

465 CT scans from collaborating institutions. We resample the CT images to isotropic spacing of 1.0 mm and crop or pad evenly along each dimension to obtain 256 × 256 × 256 volumes with the spine ROI approximately in the center. We select 418 scans for training and validation, and 47 scans are used for testing. We simulate X-rays with resolution of 0.798 mm × 0.798 mm, and size of 256 × 256. For testing, we use 500 simulated X-ray images with angles of N(0, 20) degrees in three directions, with translation in mm of N(0, 30) for in-plane (X and Y) direction and N(0, 60) for depth (Z) direction.

• Evaluation Metrics

- Mean Target Registration Error (mTRE). This metric computes the mean distance of corresponding landmarks between the warped and the target image. The mTREs re-ported in our experiments are the top 50%, 75%, and 95%(in millimeters) of the synthesis images.
- Success Rate (SR). In addition, we also report the success rate of the registration, which is defined as the percentage of the tested cases with a TRE smaller than 10 mm

• Comparison with Existing Methods

As shown in Table 1, our method outperforms existing fully-differentiable methods in terms of the top 50%, 75%, and 95% of the mTRE, demonstrating superior perfor-mance. Additionally, our approach exhibits a higher success rate throughout the ex-periment, indicating its robustness. Furthermore, it suggests a broader capture range compared to existing methods. Moreover, we provide several qualitative results of our proposed registration method in Fig. 2. The robust performance of this method demon-strates its strong controllability.

Method	mTRE(mm)↓			SR(%)↑
	Top 95%	Top 75%	Top 50%	
Initial	225.7± 89.7	188.6±64.5	148.0±47.3	
+CMA-ES	98.6±98.6	55.7±49.9	24.2±14.7	22.0
ProST	185.7± 90.2	151.2± 65.8	114.3±46.3	
+CMA-ES	38.3±55.5	12.8±19.7	2.6±1.3	55.6
SOPI	163.6±78.9	133.1±56.9	101.3±39.9	
+CMA-ES	34.7±55.8	9.1 ±13.7	2.2±1.0	58.4
Ours	155.7±69.6	127.2±43.8	95.1±30.1	
+CMA-ES	32.0±51.9	7.9±11.9	2.2±0.9	**61.0**

Table 1: 2D/3D registration performance comparing with the baseline methods. This evaluation includes measurement of the mean Target Registration Error (mTRE) at top 50%, 75%, and 95%, as well as calculating the success rate (SR) of registration.

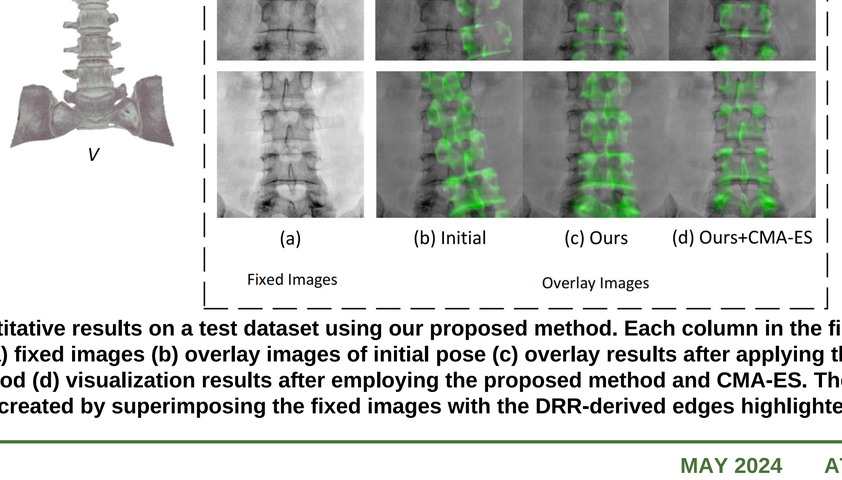

Fig. 2 Quantitative results on a test dataset using our proposed method. Each column in the figures represents: (a) fixed images (b) overlay images of initial pose (c) overlay results after applying the proposed method (d) visualization results after employing the proposed method and CMA-ES. The overlay images are created by superimposing the fixed images with the DRR-derived edges highlighted in green.