# FIDELITY-IMPOSED DISPLACEMENT EDITING FOR THE LEARN2REG 2024 SHG-BF CHALLENGE

*Jiacheng Wang*[*]   *Xiang Chen*[†]   *Renjiu Hu*[‡]   *Rongguang Wang*[§]   *Min Liu*[†]
*Yaonan Wang*[†]   *Jiazheng Wang*[†]   *Hao Li*[*]   *Hang Zhang*[‡]

[*] Vanderbilt University, Nashville, USA
[†] Hunan University, Hunan, China
[‡] Cornell University, Ithaca, USA
[§] University of Pennsylvania, Philadelphia, USA

## ABSTRACT

Co-examination of second-harmonic generation (SHG) and bright-field (BF) microscopy enables the differentiation of tissue components and collagen fibers, aiding the analysis of human breast and pancreatic cancer tissues. However, large discrepancies between SHG and BF images pose challenges for current learning-based registration models in aligning SHG to BF. In this paper, we propose a novel multi-modal registration framework that employs fidelity-imposed displacement editing to address these challenges. The framework integrates batch-wise contrastive learning, feature-based pre-alignment, and instance-level optimization. Experimental results from the Learn2Reg COMULISglobe SHG-BF Challenge validate the effectiveness of our method, securing the 1st place on the online leaderboard.

***Index Terms*—** Second-harmonic generation, Image registration, Contrastive learning.

## 1. INTRODUCTION

Image registration is a fundamental task in medical imaging, crucial for aligning images from different modalities or time points. Second-harmonic generation (SHG) microscopy provides high-resolution images sensitive to collagen fibers, while bright-field (BF) microscopy with hematoxylin and eosin (H&E) staining highlights various tissue components [1, 2]. Accurate registration of SHG and BF images is essential for comprehensive cancer tissue analysis, offering deeper insights into tissue structure and pathology.

SHG-BF registration presents two primary challenges. First, the two modalities differ significantly: SHG images emphasize collagen fibers, while BF images highlight stained tissue components, resulting in large visual discrepancies. Second, the sparse distribution of highlighted structures in SHG creates a severe foreground-background imbalance, making it difficult for learning-based methods, including convolutional neural networks [3, 4, 5, 6], vision transformers

[7, 8], and keypoint-based methods [9], to perform effectively.

To address these challenges, we propose a novel SHG-BF multimodal registration method with the following key contributions:
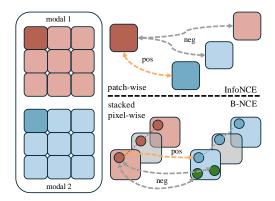
1. **Batch-wise contrastive loss (B-NCE)**: We introduce a batch-wise noise contrastive estimation loss to effectively capture shared features between SHG and BF images.
2. **Feature-based prealignment and instance optimization**: A prealignment step using descriptor matching is followed by instance-level optimization to refine the registration.
3. **Flexible transition of similarity metrics**: We combine local normalized cross-correlation (LNCC) and cross mutual information function (CMIF) as similarity metrics, balancing global and local alignment.

The novel contrastive learning loss addresses the modality gap, while the instance optimization overcomes the foreground-background imbalance. Quantitative and qualitative results demonstrate considerable improvements in registration accuracy and robustness, earning us 1st place on the online leaderboard of the Learn2Reg COMULISglobe SHG-BF Challenge.
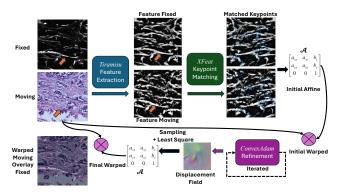
## 2. RELATED WORK

**Classical methods.** Traditional registration methods typically formulate the problem as an iterative energy optimization, employing similarity criteria such as LNCC, mean-squared error (MSE), or mutual information (MI). Approaches like ANTS [10] and ELASTIX [11] utilize gradient descent for optimization but are prone to local minima, particularly when displacements exceed feature scales. Discrete optimization techniques based on Markov Random Fields (MRF) [12] and local cost aggregation [13, 14] have been proposed to mitigate these issues.

**Regression-Based Learning.** Regression-based methods [15] train convolutional neural networks (ConvNets) to directly estimate the affine matrix. However, ConvNets lack inherent coordinate information, making affine matrix regres-

**Fig. 1**: Illustration of our proposed batch-wise noise contrastive estimation (B-NCE) loss, which aggregates pixel-level information across patches, compared to traditional InfoNCE loss that operates directly on patch-level information.



**Fig. 2**: Overview of our proposed framework, which follows three main steps: feature extraction, XFeat feature matching, and ConvexAdam fine-tuning.

sion challenging. Vision transformers [7] and keypoint-based methods like KeyMorph [9] attempt to overcome these limitations by leveraging invariant feature representations, but still struggle with the high foreground-background imbalance.

**Descriptor Matching Learning.** Descriptor matching methods [16, 17, 18, 19] effectively address the high foreground-background imbalance by relying on feature descriptors and feature matching, utilizing both handcrafted and learning-based descriptors. Traditional RANSAC [20] and modern graph- and attention-based approaches [21] are commonly used for matching. However, these approaches lack a fidelity loss, leaving the output predictions unconstrained.

## 3. METHODS

Our proposed registration framework, illustrated in Fig. 2, consists of two main stages: a feature-based prealignment using contrastive representation learning and a test-time instance-level optimization for final registration. Initially, we learn a shared representation between SHG and BF images using our specially designed Batch-wise Noise Contrastive

Estimation (B-NCE) loss. This shared latent space simplifies the multimodal registration problem into a monomodal one, making alignment more tractable. During inference, we refine the registration using ConvexAdam optimization, leveraging both cross-mutual information function (CMIF) and normalized cross-correlation (LNCC) as similarity metrics to achieve accurate alignment.

Let $I_A, I_B \in \mathbb{R}^2$ denote the fixed and moving images, respectively. Our goal is to construct affine transformations $\Phi(I_A, I_B)$ for each image pair, mapping $\mathbb{R}^2 \rightarrow \mathbb{R}^2$. We aim to align the transformed moving image to the fixed image, such that $\Phi(I_A, I_B) \odot I_B \sim I_A$ where $\odot$ denotes the spatial transformation of $I_B$ using $\Phi(I_A, I_B)$.

### 3.1. Features prealignment with contrastive learning

Due to the significant differences between SHG and BF images, direct registration is challenging. To address this, we map both images into a common latent space where shared features can be effectively captured. We achieve this using contrastive representation learning with a specially designed Batch-wise Noise Contrastive Estimation (B-NCE) loss.

Inspired by CoMIR [22], we employ two identical Dense Tiramisu networks [23], $D_\theta$, to extract feature representations from $I_A$ and $I_B$: $C_A = D_\theta(I_A)$ and $C_B = D_\theta(I_B)$, where $C_A, C_B \in \mathbb{R}^{H \times W \times C}$ and $C$ is the number of feature channels. And we demonstrated the trained features shown in Fig. 3 second row.

**Batch-wise noise contrastive estimation (B-NCE) loss.** Our B-NCE loss is designed to enhance the network's ability to capture shared features between modalities by focusing on position-level similarities across the batch dimension. We extract patches from $C_A$ and $C_B$ of size $p \times p$, resulting in sets of patches $\{X_k\}$ and $\{Y_k\}$ for $k = 1, \ldots, N$, where $N$ is the number of patches. To compute the loss, we consider the features at corresponding spatial locations across the batch:

$$\mathbf{x}_{i,j} = \{X_k(i,j)\}_{k=1}^N, \quad \mathbf{y}_{i,j} = \{Y_k(i,j)\}_{k=1}^N, \quad (1)$$

where $(i, j)$ indexes the spatial positions within each patch.

We define the similarity between features using the cosine similarity:

$$s(\mathbf{x}_{i,j}, \mathbf{y}_{i,j}) = \frac{\mathbf{x}_{i,j}^\top \mathbf{y}_{i,j}}{\|\mathbf{x}_{i,j}\| \|\mathbf{y}_{i,j}\|}. \quad (2)$$
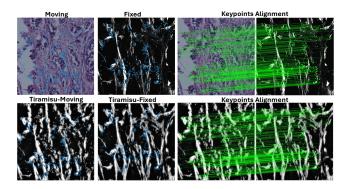
The B-NCE loss for each spatial position is:

$$\mathcal{L}_{\text{B-NCE}}^{i,j} = -\log \frac{\exp(s(\mathbf{x}_{i,j}, \mathbf{y}_{i,j})/\tau)}{\sum_{k=1}^N \exp(s(\mathbf{x}_{i,j}, \mathbf{z}_k)/\tau)}, \quad (3)$$

where $\mathbf{z}_k$ includes both positive and negative samples, and $\tau$ is a temperature parameter.

The total loss is averaged over all spatial positions:

$$\mathcal{L}_{\text{B-NCE}} = \frac{1}{H_p W_p} \sum_{i=1}^{H_p} \sum_{j=1}^{W_p} \mathcal{L}_{\text{B-NCE}}^{i,j}, \quad (4)$$

**Fig. 3**: XFeat feature detection/description/matching in SHG/BF images and the Tiramisu feature images.



**Fig. 4**: Qualitative results on the validation set. From left to right: warped BF image overlaid on SHG image, SHG image (fixed) shown in Virdis colormap, BF image (moving), and warped BF image.

where $H_p$ and $W_p$ are the height and width of the patches.

By minimizing $\mathcal{L}_{\text{B-NCE}}$, the network learns to bring corresponding features from $I_A$ and $I_B$ closer in the latent space while pushing apart non-corresponding features, effectively capturing shared structures across modalities.

**Feature-based prealignment.** With the learned representations $C_A$ and $C_B$, we perform feature-based prealignment. We detect keypoints and extract descriptors using a method inspired by XFeat [24], adapted to our context. We match features between $C_A$ and $C_B$ to compute an initial affine transformation $\Phi_{\text{init}}$. We also demonstrated the matching keypoints in Fig. 3 in blue dots and connected with green lines.
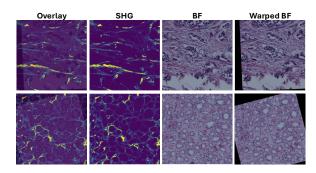
This prealignment step reduces large discrepancies between the images, providing a good starting point for the subsequent optimization.

### 3.2. Test-time Instance Optimization

We adopt the core principles of the ConvexAdam method [25], enabling flexible instance optimization between affine and deformable transformations. Unlike the original Convex-Adam, which optimizes flow parameters using hand-crafted features, our approach leverages learned features from the preceding stage and employs a pyramid structure for multi-resolution alignment across scales $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$. This hierarchical setup allows for efficient alignment at both large and fine scales.

Following the prealignment performed during the training stage, we use a convolutional network with large kernel sizes to generate the displacement fields. Displacement assignment is achieved by evaluating patch similarity costs and identifying the minimal cost positions. Smooth regularization is then applied through Gaussian smoothing, as described in [26]. Our fidelity loss combines Local Normalized Cross-Correlation (LNCC), which handles local similarities to address the foreground-background imbalance, with Cross Mutual Information Function (CMIF), which considers global intensity relationships to bridge the modality gap.

Since the primary features in SHG images are sparsely distributed, we avoid using the deformation field directly.

Instead, we sample displacements based on SHG intensity, followed by least-squares analysis on these sparse displacements. This results in an over-determined system of equations that we solve to obtain the final affine transformation. This approach allows for a smooth transition between deformable and affine transformations, enhancing the precision of the registration results.

### 3.3. Multimodal Similarities: Mutual Information

We compute mutual information between fixed image $I_A \subseteq X_A$ and moving image $I_B \subseteq X_B$ based on cross-mutual information function (CMIF) [27, 28]. Given images $I_A$ and $I_B$ intersecting on $I_{AB}$ and $I_{AB} \neq \emptyset$, we firstly cluster each image $I_A$ and $I_B$ into level sets $\mathcal{A} \in \mathbb{Z}$ and $\mathcal{B} \in \mathbb{Z}$ accordingly using K-means. For each level $a$ and $b$ from the level sets $\mathcal{A}$ and $\mathcal{B}$, we compute the marginal and joint histogram entries separately, Then, based on the normalized histograms, we derive the marginal and joint Shannon entropies as follows,

$$H_A = -\sum_{a \in \mathcal{A}} \frac{I_A^a \cdot I_B}{N_{AB}} \log \frac{I_A^a \cdot I_B}{N_{AB}}; \quad (5)$$

$$H_B = -\sum_{b \in \mathcal{B}} \frac{I_A \cdot I_B^b}{N_{AB}} \log \frac{I_A \cdot I_B^b}{N_{AB}}; \quad (6)$$

$$H_{AB} = -\sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \frac{I_A^a \cdot I_B^b}{N_{AB}} \log \frac{I_A^a \cdot I_B^b}{N_{AB}}, \quad (7)$$

where $I_A^a$ and $I_B^b$ represent images where the pixel values equal to level set value $a$ and $b$ respectively; $N_{AB}$ is the number of total pixels in both image $I_A$ and $I_B$. Finally, we can compute the mutual information using $MI(I_A, I_B) = H_A + H_B - H_{AB}$.

**Table 1**: TRE comparison of teams on the Learn2Reg 2024 Challenge validation leaderboard. Rankings are based on a snapshot of the leaderboard as of Sep 20, 2024, excluding challenge organizer entries.

| Method | TRE (LM) | Best Validation Place |
|---|---|---|
| Team VROC | $2.620 \pm 1.206$ | 36th |
| Team Yangzhao | $2.578 \pm 1.377$ | 34th |
| Team IWM | $2.077 \pm 1.182$ | 8th |
| Team meeem | $\underline{2.032 \pm 1.026}$ | 6th |
| XFeat Only | $5.939 \pm 8.198$ | 39th |
| XFeat + Ours (CMIF) | $9.128 \pm 1.211$ | 72th |
| XFeat + Ours (LNCC) | $2.361 \pm 0.812$ | 20th |
| XFeat + Ours (LNCCC+CMIF) | $\mathbf{1.943 \pm 0.765}$ | 1st |

## 4. EXPERIMENTS AND RESULTS

### 4.1. Dataset

We evaluated our method on the dataset from the Learn2Reg 2024 Challenge Task 3: COMULISglobe SHG-BF [1], which includes 156 training, 10 validation, and 40 test cases of SHG and H&E-stained BF microscopy images of breast and pancreatic cancer tissues. Acquired at the University of Wisconsin-Madison, these images present significant challenges due to differences in imaging modalities: SHG images emphasize collagen fibers, while BF images highlight H&E-stained tissue structures. Reliable expert annotations provide landmarks for validation and testing.
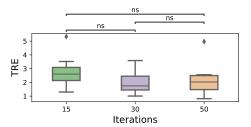
### 4.2. Data Preprocessing and Evaluation Metrics

For preprocessing, we applied stochastic intensity transformations using Bézier curves [29] and random affine augmentations, including scaling, rotation, and shearing, to enhance the robustness of our model to intensity variations and geometric distortions. During validation and testing, we omitted stochastic augmentations but maintained normalization and intensity adjustments to ensure consistency.

We adopted the evaluation metrics established by the Learn2Reg 2024 Challenge, specifically using Target Registration Error (TRE) as the primary metric to assess registration accuracy. TRE measures the average Euclidean distance between corresponding anatomical landmarks in the fixed and moving images after registration. Lower TRE values indicate better alignment. All reported results on the validation set are based on the challenge leaderboard.

### 4.3. Quantitative Results

Table 1 presents the performance of our method compared to top-performing methods from the challenge. Our approach achieved a mean TRE of 1.943 mm on the validation set, ranking first. Notably, our method demonstrated the lowest standard deviation (0.765 mm) among all submissions, indicating high robustness and consistency across different cases.



**Fig. 5**: Impact of the number of optimization iterations on the mean TRE. Our method achieves consistent performance, with the best result at 30 iterations.

To evaluate the effectiveness of our proposed components, we conducted ablation studies summarized in Table 1 (second part). Introducing instance-level optimization using LNCC reduced the TRE by approximately 60%, effectively mitigating the imbalance caused by sparse SHG features. Using only CMIF as the optimization metric increased the TRE, suggesting that CMIF alone may lead to local minima due to its global nature and the modality gap. Combining CMIF with LNCC in our fidelity loss improved performance, reducing the TRE from 2.361 to 1.943. This demonstrates that our fidelity-imposed displacement editing, which balances global and local alignment, enhances registration accuracy.

Additionally, we tested different iteration counts during instance-level optimization (15, 30, and 50 iterations) using both CMIF and LNCC (Fig. 5). Our method consistently achieved low TRE values, with the best performance at 30 iterations (mean TREs of 2.759, 1.943, and 2.168, respectively). This indicates our approach's stability without extensive computation.

### 4.4. Qualitative Results

Two validation examples are presented in Fig. 4. Despite significant structural differences due to the modality gap, our method successfully aligned the moving BF images to the fixed SHG images. The overlaid warped BF images show main structures aligned with minimal misalignment, demonstrating effective multimodal registration.

Figure 3 shows feature matching with our learned representations. Despite large modality discrepancies, the features are visually similar, enabling robust matching. The sufficient matching pairs confirm the effectiveness of B-NCE loss in bridging the modality gap for feature-based prealignment.

## 5. CONCLUSION

In this paper, we propose a new affine registration framework that integrates a novel contrastive learning loss (B-NCE) with rich contextual information from modern neural networks inside fidelity-imposed iterative instance optimization. Our fidelity-imposed network editing framework overcomes modality gap between SHG and BF images, providing robust

and accurate SHG-BF registration performance. Experimental results from the Learn2Reg COMULISglobe SHG-BF Challenge demonstrate the effectiveness of our approach, earning the 1st place on the online leaderboard.

## 6. REFERENCES

[1] Kevin Eliceiri et al., "Multimodal biomedical dataset for evaluating registration methods (full-size tma cores)," *Zenodo, Feb*, 2021.

[2] Adib Keikhosravi et al., "Intensity-based registration of bright-field and second-harmonic generation images of histopathology tissue sections," *Biomedical Optics Express*, vol. 11, no. 1, pp. 160–173, 2019.

[3] Hang Zhang et al., "Memwarp: Discontinuity-preserving cardiac registration with memorized anatomical filters," in *MICCAI*. Springer, 2024, pp. 671–681.

[4] Xiang Chen et al., "Spatially covariant image registration with text prompts," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[5] Guha Balakrishnan et al., "Voxelmorph: a learning framework for deformable medical image registration," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[6] Hang Zhang et al., "Slicer networks," *arXiv preprint arXiv:2401.09833*, 2024.

[7] Tony CW Mok and Albert Chung, "Affine medical image registration with coarse-to-fine vision transformer," in *CVPR*, 2022, pp. 20835–20844.

[8] Junyu Chen et al., "Transmorph: Transformer for unsupervised medical image registration," *Medical image analysis*, vol. 82, pp. 102615, 2022.

[9] M Yu Evan et al., "Keymorph: Robust multi-modal affine registration via unsupervised keypoint detection," in *MIDL*, 2022.

[10] B B Avants et al., "Advanced normalization tools (ants)," *Insight j*, vol. 2, no. 365, pp. 1–35, 2009.

[11] Stefan Klein et al., "Elastix: a toolbox for intensity-based medical image registration," *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 196–205, 2009.

[12] Mattias P Heinrich et al., "Mrf-based deformable registration and ventilation estimation of lung ct," *IEEE transactions on medical imaging*, vol. 32, no. 7, pp. 1239–1248, 2013.

[13] Mattias P Heinrich et al., "Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities.," *ISBI*, vol. 1390, pp. 27, 2015.

[14] Frank Steinbrücker et al., "Large displacement optical flow computation without warping," in *ICCV*, 2009, pp. 1609–1614.

[15] Xu Chen et al., "Learning unsupervised parameter-specific affine transformation for medical images registration," in *MICCAI*. Springer, 2021, pp. 24–34.

[16] Daniel DeTone et al., "Superpoint: Self-supervised interest point detection and description," in *CVPR workshops*, 2018, pp. 224–236.

[17] Prune Truong et al., "Glampoints: Greedily learned accurate match points," in *ICCV*, 2019, pp. 10732–10741.

[18] Jiacheng Wang et al., "Novel oct mosaicking pipeline with feature-and pixel-based registration," in *ISBI*. IEEE, 2024, pp. 1–5.

[19] Jiacheng Wang et al., "Retinal ipa: Iterative keypoints alignment for multimodal retinal imaging," *arXiv preprint arXiv:2407.18362*, 2024.

[20] Robert C Bolles and Martin A Fischler, "A ransac-based approach to model fitting and its application to finding cylinders in range data.," in *IJCAI*, 1981, vol. 1981, pp. 637–643.

[21] Paul-Edouard Sarlin et al., "Superglue: Learning feature matching with graph neural networks," in *CVPR*, 2020, pp. 4938–4947.

[22] Nicolas Pielawski et al., "Comir: Contrastive multimodal image representation for registration," *Advances in neural information processing systems*, vol. 33, pp. 18433–18444, 2020.

[23] Simon Jégou et al., "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *CVPR workshops*, 2017, pp. 11–19.

[24] Guilherme Potje et al., "Xfeat: Accelerated features for lightweight image matching," in *CVPR*, 2024, pp. 2682–2691.

[25] Hanna Siebert et al., "Fast 3d registration with accurate optimisation and little learning for learn2reg 2021," in *MICCAI*. Springer, 2021, pp. 174–179.

[26] Tom Vercauteren et al., "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.

[27] Bernd Pompe et al., "Using mutual information to measure coupling in the cardiorespiratory system," *IEEE Engineering in Medicine and Biology Magazine*, vol. 17, no. 6, pp. 32–39, 1998.

[28] Johan Öfverstedt et al., "Fast computation of mutual information in the frequency domain with applications to global multimodal image alignment," *Pattern Recognition Letters*, vol. 159, pp. 196–203, 2022.

[29] Ken Kobayashi et al., "Bézier simplex fitting: Describing pareto fronts of simplicial problems with small samples in multi-objective optimization," in *AAAI*, 2019, vol. 33, pp. 2304–2313.