VFMStitch: A Vision-Foundation-Model Empowered Framework for 3D Ultrasound Stitching via Geometric–Semantic Feature Fusion

Xing Yao[1], **Nick DiSanto**[1], **Runxuan Yu**[1], **Jiacheng Wang**[1], **Daiwei Lu**[1],**Gabriel Arenas**[2], **Baris Oguz**[2], **Alison Pouch**[2], **Nadav Schwartz**[2], **Brett C Byram**[1], **Ipek Oguz**[1]

[1] *Vanderbilt University,* [3] *University of Pennsylvania*

Editors: Under Review for MIDL 2026

Abstract

3D ultrasound (3DUS) stitching expands the field-of-view (FOV) by registering partially overlapping 3DUS volumes acquired from different probe positions. This task is intrinsically difficult due to large inter-volume translations and rotations, the impact of the sector-shaped FOV, as well as the heavy noise and artifacts inherent to ultrasound. With the rapid progress of Vision Foundation Models (VFMs) such as DINOv3, VFM-derived features have recently shown promise for downstream medical image registration tasks. However, existing VFM-based approaches primarily focus on deformable registration and are rarely evaluated for rigid alignment under large motions. Moreover, the feasibility of leveraging VFM-derived features for robust 3DUS stitching remains largely unexplored. In this study, we introduce **VFMStitch**, the first training-free, VFM-empowered 3DUS stitching framework that integrates point-cloud (PCD)–based geometric features with DINOv3-derived semantic descriptors. Extensive experiments demonstrate that VFMStitch substantially improves rigid registration accuracy compared to existing methods, validating the effectiveness of geometric–semantic fusion for challenging 3DUS stitching scenarios. The code is available at github.com/MedICL-VU/VFMStitch.

Keywords: vision foundation model, DINOv3, ultrasound, stitching, point cloud, feature fusion

1. Introduction

Ultrasound (US) image registration (Che et al., 2017; Entrekin et al., 2001; Wang et al., 2014) is a pivotal task to many downstream analysis tasks, with US stitching (Banerjee et al., 2015; Gomez et al., 2019; Wright et al., 2023; Bano and Stoyanov, 2024) being a key application for expanding the field of view (FOV) by aligning partially overlapping scans from different probe positions. This is particularly important for visualizing large anatomical structures such as the fetus and placenta (Roy-Lacroix et al., 2017; Gomez et al., 2017). However, large inter-volume translations and rotations frequently arise during freehand scanning (Yao et al., 2025b), making rigid alignment for 3DUS stitching inherently difficult. Furthermore, the effect of the sector-shaped FOV (Yao et al., 2025a), low signal-to-noise ratio and artifacts of US imaging (Yao et al., 2024) weaken the reliability of intensity-based similarity measures and reduce the distinctiveness of local features. These factors collectively impose significant limitations on intensity-based or feature-based registration pipelines and motivate the exploration of more robust, representation-driven solutions.

To address these challenges, prior studies have explored a variety of strategies, including block-matching (Banerjee et al., 2015), manifold-learning–based keypoint selection (Gomez et al., 2019, 2017), iterative spatial transformer network (Wright et al., 2019), and reinforcement learning (Wright et al., 2023). More recently, diffusion-based frameworks such as SynStitch (Yao et al., 2025b) and LOTUS (Yao et al., 2025a) have been proposed to explicitly mitigate the impact of the sector-shaped FOV in both 2D and 3D US stitching scenarios. While these approaches represent meaningful advances, they often require large-scale training data and remain sensitive to variations in image quality. A training-free, robust, and accurate 3DUS stitching framework is thus desirable.

Recently, self-supervised vision foundation models (VFMs), such as the DINO family (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025), have attracted substantial attention due to their strong generalization ability and competitive zero-shot performance on a wide range of medical imaging tasks (Ambsdorf et al., 2025; Li et al., 2025; Avants et al., 2008; Wang et al., 2025; Gu et al., 2025). Among these efforts, DINO-Reg (Avants et al., 2008) and DINOv3+T^3 (Wang et al., 2025) have demonstrated state-of-the-art results in deformable medical image registration by leveraging DINOv2 (Oquab et al., 2023) and DINOv3 (Siméoni et al., 2025) features in combination with test-time optimization strategies (Siebert et al., 2024). Despite this promising progress, significant gaps remain. First, existing VFM-based registration frameworks primarily target deformable alignment, leaving rigid registration under large translations and rotations insufficiently explored. Second, prior evaluations have been conducted largely on high-quality modalities such as MRI and CT, where texture and contrast are substantially more stable. Their performance in artifact-prone, and highly heterogeneous 3D ultrasound, particularly in demanding scenarios such as placenta imaging, remains an open question. These limitations motivate the exploration of how VFMs can be adapted or extended to address the unique difficulties of 3DUS stitching.

In this work, we propose **VFMStitch**, the first training-free, VFM-empowered framework specifically designed for 3DUS stitching under large rigid motions by performing point-cloud (PCD)–based registration on the fusion of PCD-based geometric features with DINOv3-derived semantic descriptors. Both qualitative and quantitative evaluations show that VFMStitch significantly outperforms state-of-the-art methods in challenging 3DUS stitching scenarios. Our main contributions are as follows:

- **Methodological novelty:** VFMStitch is the first framework to leverage VFM-derived features for medical image registration involving *large rigid transformations*, filling a critical gap left by existing deformable-focused VFM approaches.
- **Geometric robustness:** We show that PCD-based registration serves as a reliable and robust alternative to intensity-based registration for 3DUS stitching, particularly in the presence of strong noise and artifacts.
- **Semantic descriptor superiority:** We demonstrate that DINOv3-based semantic features significantly outperform traditional and learning-based descriptors when used as PCD registration features.
- **Effectiveness of geometric-semantic feature fusion:** Our results further reveal that fusing geometric and semantic cues yields additional performance gains, highlighting the complementary nature of these representations.

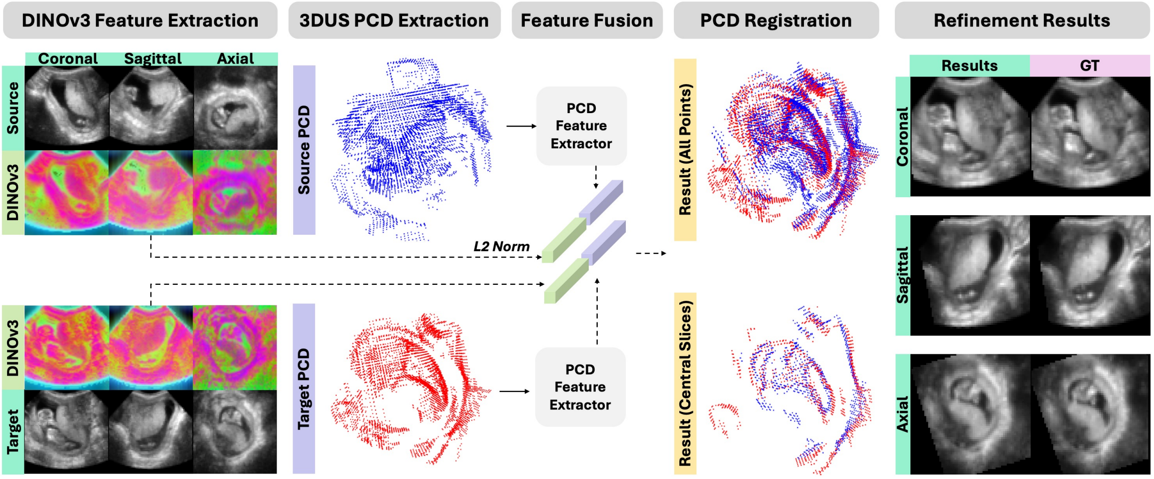

Figure 1: Pipeline of VFMStitch. DINOv3-derived semantic features are extracted from the stitching pair, followed by 3DUS PCD extraction and geometric descriptor computation. Semantic and geometric features are fused for PCD-based registration, and the result is refined via unsupervised intensity-based registration.

2. Methods

Fig. 1 illustrates the VFMStitch pipeline. A stitching pair $(I_{\text{source}}, I_{\text{target}})$ refers to two partially overlapping 3DUS volumes acquired from the same subject at different probe positions. For each stitching pair, DINOv3-derived semantic features are first extracted, with the top three principal components visualized in RGB via principal component analysis (PCA).Point clouds (PCDs) are then extracted from the corresponding gray-scale volumes, and geometric descriptors are computed. The semantic and geometric features are fused to perform PCD-based registration from the source to target image. Finally, the estimated transformation is propagated to voxel space and refined through unsupervised intensity-based registration.

2.1. Unsupervised 3DUS PCD Extraction

As intensity-based registration struggles under heterogeneous image quality, geometric alternatives based on PCDs have been explored. In ultrasound applications, PCDs are often extracted from segmentation masks (Jiang et al., 2023, 2024; Tan et al., 2023), which is suboptimal for 3DUS stitching: manual annotation is costly, automated segmentation generalizes poorly, and object-centric masks may discard contextual anatomy when inter-volume overlap is limited, resulting in poor registration.

To address these limitations, we propose an unsupervised, segmentation-free PCD extraction strategy based on edge representations that preserve rich anatomical geometry. Specifically, edge maps are computed using complementary classical operators, including the Sobel filter (Kanopoulos et al., 1988), Laplacian of Gaussian (LoG) (Marr and Hildreth, 1980), and Harris corner detector (Harris et al., 1988), capturing intensity gradients, fine structural ridges, and stable corner-like features, respectively. Field-of-view (FOV) masks are applied to suppress boundary artifacts induced by the sector-shaped 3DUS geometry, retaining only anatomically relevant regions. The filtered edge maps are then averaged and normalized to the range $[0, 1]$ to form a unified geometric representation. Compared to intensity- or segmentation-based cues, this edge-based representation provides a more complete and noise-robust characterization of 3DUS anatomy for registration initialization.

For a given 3DUS volume $I \in \mathbb{R}^{D\times H\times M}$, the final 3DUS point cloud is denoted as $\{\mathbf{p}_i\}_{i=1}^N$, where $\mathbf{p}_i \in \mathbb{R}^3$. $\{\mathbf{p}_i\}_{i=1}^N$ is extracted by thresholding the merged edge map and subsequently the number of points is reduced by 1/5 using Open3D (Zhou et al., 2018), reducing computational complexity while preserving geometric fidelity.

2.2. Geometric-Semantic Feature Fusion

PCD-based geometric descriptor extraction: To extract a local geometric descriptor $\mathbf{f}^{\text{Geo}}i \in \mathbb{R}^{d\text{Geo}}$ for each point $\mathbf{p}_i$, we investigate a diverse set of PCD geometric descriptors, including both conventional hand-crafted methods and state-of-the-art learning-based approaches. Specifically, we evaluate Fast Point Feature Histograms (FPFH) (Rusu et al., 2009) as a classical baseline, as well as recent learning-based models, including Point-MAE (Pang et al., 2022), Point-DAE (Zhang et al., 2025), and Point-BERT (Yu et al., 2022). For FPFH, normal vectors are estimated via KD-tree search (radius: 10, up to 30 neighbors), followed by descriptor extraction with a larger neighborhood (radius: 25, up to 100 neighbors), yielding features of dimensionality $d_{\text{Geo}} = 33$. Learning-based descriptors are extracted using pretrained models without fine-tuning, producing per-point features with a unified dimensionality of $d_{\text{Geo}} = 384$ to ensure fair comparison.

DINOv3-based semantic feature extraction: Each 3DUS volume $I \in \mathbb{R}^{D\times H\times M}$ (with $D = H = M = 64$) is split into its $H \times M$ slices along the sagittal axis, and DINOv3 features are extracted on a per-slice basis. Voxel intensities are linearly normalized to $[0, 255]$, and each single-channel slice is replicated across three channels to form pseudo-RGB inputs. The slices are upsampled by a factor of $s = 16$, corresponding to the ViT patch size, to ensure spatial alignment between the output feature maps and the input images. The resulting inputs are standardized using ImageNet mean and standard deviation and fed into a pretrained DINOv3 ViT-L encoder. Finally, patch tokens from the last layer are reshaped to produce feature maps $F_{\text{Dense}} \in \mathbb{R}^{H\times M\times C}$ for each slice, where $C = 1024$.

For each stitching pair $(I_{\text{source}}, I_{\text{target}})$, dense feature maps from all slices of both volumes are aggregated and projected into a shared low-dimensional embedding space via PCA. This yields PCA-compressed feature maps $F_{\text{PCA}} \in \mathbb{R}^{H\times M\times d_{\text{DINO}}}$ for each 2D slice, where $d_{\text{DINO}} = 16$. For each volume, the compressed feature maps are then stacked along the slice dimension to form a 4D feature volume $V \in \mathbb{R}^{D\times H\times M\times d_{\text{DINO}}}$, which is spatially aligned with the original voxel grid for subsequent registration and analysis.

Mapping DINOv3 features to PCD: Next, DINOv3 descriptor $\mathbf{f}_i^{\text{DINO}} \in \mathbb{R}^{d_{\text{DINO}}}$ for each $\mathbf{p}_i$ are extracted by sampling V in the voxel space.

We perform trilinear interpolation over the eight neighboring grid points of p_i, which yields $\mathbf{f}_i^{\text{DINO}}$ for each $\mathbf{p}_i$. Collecting all descriptors, we obtain a DINOv3 descriptor set $F_{\text{DINO}} \in \mathbb{R}^{N\times d_{\text{DINO}}}$ that mapping the DINOv3 feature V from voxel space to PCD space.

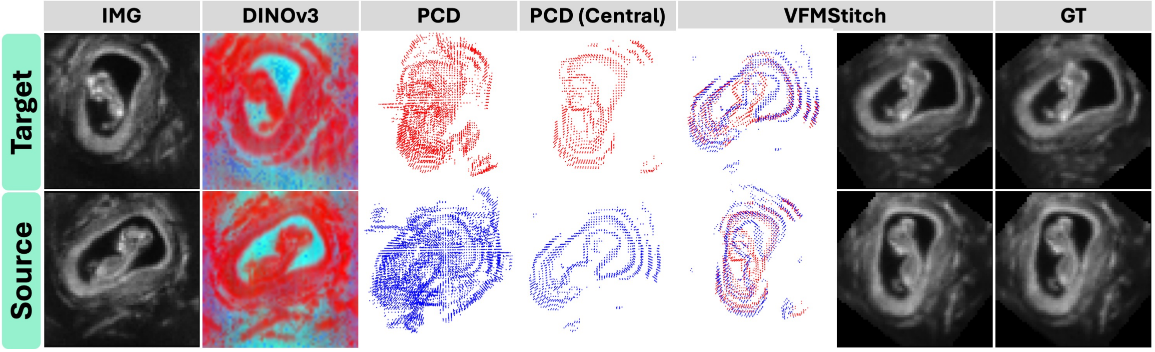

Figure 2: Visualization of the intermediate steps of VFMStitch in the axial plane. The "PCD (Central)" panel shows the extracted point clouds from the central axial slice only for clarity. The first row illustrates registration from target to source, while the second row shows registration from source to target, demonstrating the bidirectional alignment capability of VFMStitch.

Geometric-semantic feature fusion: For each downsampled point $\mathbf{p}_i$, we fuse its local geometric descriptor $\mathbf{f}_i^{\text{Geo}}$ with the corresponding semantic descriptor $\mathbf{f}_i^{\text{DINO}}$. We first apply row-wise ℓ_2-normalization to each modality,

$$\tilde{\mathbf{f}}_i^{\text{Geo}} = \frac{\mathbf{f}_i^{\text{Geo}}}{\|\mathbf{f}_i^{\text{Geo}}\|_2 + \varepsilon}, \quad \tilde{\mathbf{f}}_i^{\text{DINO}} = \frac{\mathbf{f}_i^{\text{DINO}}}{\|\mathbf{f}_i^{\text{DINO}}\|_2 + \varepsilon},$$

with a small constant $\varepsilon = 10^{-8}$ for numerical stability. We then form a weighted concatenation of geometric and DINO descriptors,

$$\mathbf{f}_i^{\text{fuse}} = \left[\alpha\tilde{\mathbf{f}}_i^{\text{Geo}}\,;\, \beta\tilde{\mathbf{f}}_i^{\text{DINO}}\right] \in \mathbb{R}^{d_{\text{Geo}}+d_{\text{DINO}}},$$

where $\alpha, \beta > 0$ control the relative contributions of geometric structure and semantic appearance (set to $\alpha = \beta = 1.0$ by default). Finally, we perform a global ℓ_2-normalization on the fused descriptors,

$$\hat{\mathbf{f}}_i^{\text{fuse}} = \frac{\mathbf{f}_i^{\text{fuse}}}{\|\mathbf{f}_i^{\text{fuse}}\|_2 + \varepsilon}.$$

These normalized fused descriptors combining complementary geometric and semantic are then used as input to the PCD-based registration.

2.3. Registration

PCD-based registration: We adopt two widely used robust PCD registration algorithms, Random Sample Consensus (RANSAC) (Fischler and Bolles, 1981) and TEASER++ (Yang et al., 2020), to estimate rigid transformations between 3DUS point clouds.

For RANSAC-based registration, an initial global alignment is obtained through descriptor-based feature matching, with geometric consistency enforced using both edge-length preservation (threshold: 0.9) and Euclidean distance constraints (threshold: 10.0). RANSAC is executed for up to 4×10^6 iterations, with early termination triggered after 500 inlier correspondences are identified. To further improve local alignment accuracy, the resulting transformation is refined using Iterative Closest Point (ICP) (Besl and McKay, 1992), which minimizes the point-to-point distances between the aligned point clouds.

For TEASER++, we perform robust global rigid registration by directly estimating the rotation $\mathbf{R} \in \mathrm{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$ between the source and target point clouds. Throughout this work, we denote the instantiations of the proposed framework using RANSAC and TEASER++ as VFMStitch-R and VFMStitch-T, respectively.

Transformation propagation to volumetric space: The resulting rigid transformation is represented as a homogeneous matrix $\mathbf{T} \in \mathbb{R}^{4\times4}$ and subsequently applied to the corresponding gray-scale 3DUS volumes. Fig. 2 visualizes the intermediate steps of VFMStitch for a randomly selected example.

Intensity-based refinement: We observe that the strong rigid initialization provided by VFMStitch facilitates subsequent intensity-based optimization, leading to further improvements in pixel-level alignment accuracy. To this end, we apply an intensity-based rigid refinement step using ANTs after VFMStitch. This refinement preserves the global alignment established by VFMStitch while exploiting local intensity consistency to further enhance registration accuracy.

2.4. Datasets and Evaluation Metrics

Dataset: We evaluate the proposed method on the in-house *RegUS* dataset, which comprises 3DUS placenta scans from 20 first-trimester pregnancies. For each subject, two partially overlapping 3DUS volumes are acquired from different probe positions. All volumes are resampled to an isotropic spatial resolution of $(2\ \mathrm{mm})^3$, centrally cropped to 64^3 voxels, and intensity-normalized to the range $[0, 1]$.

Ground truth: Ground-truth rigid transformations are manually created by two experienced experts and further visually inspected by three additional reviewers, serving as the reference standard for quantitative evaluation. The manual transformations involve large relative motions, with rotations ranging from 30° to 117° and translations spanning $[25, 83]$ mm, reflecting realistic and challenging clinical stitching scenarios. Each subject's volumes are registered bidirectionally, resulting in a total of 40 registration pairs.

Evaluation metrics: Registration performance is quantitatively assessed using standard image similarity metrics, including normalized cross-correlation (NCC) (Lewis, 1995), structural similarity index measure (SSIM) (Wang et al., 2004), peak signal-to-noise ratio (PSNR) (Gonzalez and Woods, 2008), and mean squared error (MSE).

Implementation details: We comprehensively evaluate the proposed VFMStitch by comparing it with a diverse set of representative baselines, covering conventional (ANTs (Avants et al., 2008), SIFT3D (Rister et al., 2017)), outpainting-based (LOTUS(Yao et al., 2025a)) and learning-based baselines (ConvexAdam(Siebert et al., 2024), DINOv3+T^3(Wang et al., 2025)). For ANTs, we perform rigid registration on both gray-scale 3DUS volumes (denoted as ANTs), and the extracted 4D DINOv3 feature maps (denoted as ANTs-DINO). For LOTUS, ANTs is employed as the registration method following the outpainting. ConvexAdam is evaluated under rigid registration settings using MIND (Heinrich et al., 2012) descriptors. Similarly, DINOv3+T^3 is also evaluated in rigid configurations to ensure a fair

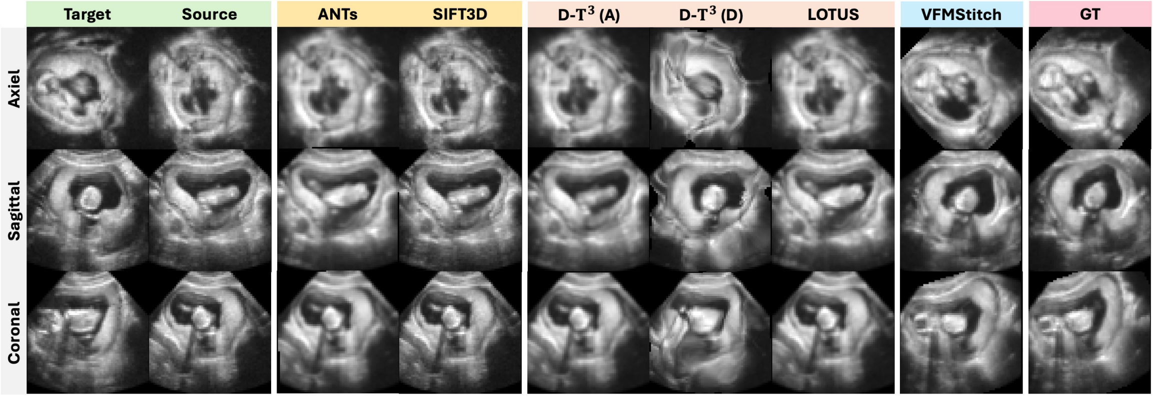

Figure 3: Qualitative comparison on a randomly selected case. D-T^3(A) and D-T^3(D) denote the rigid and deformable variants of DINOv3-T^3, respectively. While the baseline methods fail to achieve satisfactory alignment under large inter-volume motion, VFMStitch produces accurate registration. Notably, D-T^3(D) introduces unrealistic geometric distortions, as expected, highlighting the limitations of existing VFM-based deformable registration frameworks in handling large-motion 3DUS stitching.

assessment of its performance under large-motion stitching scenarios, and we also demonstrate its deformable (original setting) registration performance qualitatively. We expect that deformable registration is not suitable to capture rigid transformation and will introduce distortions. All baseline methods are configured according to their recommended settings, and no additional fine-tuning is performed beyond what is required for fair comparison.

3. Results and Discussion

Effectiveness of VFMStitch: Fig. 3 presents qualitative registration results on a randomly selected example. Conventional methods and learning-based baselines consistently fail due to the strong local minima induced by the sector-shaped FOV. In contrast, VFMStitch achieves accurate and visually coherent rigid alignment, effectively preserving anatomical structures across stitched volumes.

Notably, the results further reveal that existing VFM-based registration methods are inadequate for 3DUS stitching when substantial motion is present. Specifically, deformable registration using DINOv3-T^3 (D-T^3(D)) relies on dense deformation fields to establish pixel-wise correspondences, resulting in unrealistic anatomical distortions, as can be expected. Conversely, the rigid variant (D-T^3(A)) fails to recover the correct rigid alignment, highlighting the inherent difficulty of estimating large rotations and translations within a deformation-field optimization framework. These observations indicate a fundamental mismatch between the current VFM-based registration methods and the large motions often present in 3DUS stitching tasks.

Table 1: Comparison of VFMStitch and baseline methods on the RegUS dataset. R: RANSAC matcher; T: TEASER++ matcher; F: ANTs-based refinement. Gray: conventional methods; blue: learning-based methods; green: VFMStitch (before refinement); yellow: VFMStitch (after refinement). **Best** and second best results are highlighted. VFMStitch consistently outperforms all baselines across all metrics, with ANTs refinement providing additional gains.

Methods	NCC ↑	MSE(×10) ↓	PSNR ↑	SSIM ↑
ANTs	0.826 ± 0.107	0.174 ± 0.096	19.894 ± 6.673	0.496 ± 0.219
ANTs-DINOv3	0.767 ± 0.090	0.224 ± 0.069	16.722 ± 1.400	0.388 ± 0.072
SIFT3D	0.757 ± 0.124	0.219 ± 0.074	16.893 ± 1.736	0.403 ± 0.126
ConvexAdam	0.773 ± 0.092	0.218 ± 0.064	16.813 ± 1.368	0.391 ± 0.078
DINOv3+T^3	0.778 ± 0.083	0.217 ± 0.064	16.829 ± 1.362	0.390 ± 0.076
LOTUS	0.881 ± 0.132	0.120 ± 0.110	21.783 ± 5.140	0.609 ± 0.251
VFMStitch(DINOv3, R)	0.885 ± 0.127	0.116 ± 0.095	20.923 ± 3.957	0.598 ± 0.198
VFMStitch(DINOv3, T)	0.874 ± 0.144	0.116 ± 0.097	21.221 ± 4.390	0.610 ± 0.209
VFMStitch(DINOv3, R, F)	$\mathbf{0.910 \pm 0.114}$	$\mathbf{0.090 \pm 0.097}$	$\mathbf{24.064 \pm 6.542}$	$\mathbf{0.690 \pm 0.242}$
VFMStitch(DINOv3, T, F)	$\underline{0.891 \pm 0.135}$	$\underline{0.098 \pm 0.094}$	$\underline{23.260 \pm 6.179}$	$\underline{0.666 \pm 0.240}$

Table 1 reports a comprehensive quantitative comparison across all methods. Among existing baselines, LOTUS achieves the strongest overall performance, serving as a competitive state-of-the-art reference. However, the proposed VFMStitch with DINOv3 as descriptor consistently outperforms LOTUS across all evaluation metrics under different PCD registration strategies (RANSAC and TEASER++), with the sole exception of a slightly lower NCC score for VFMStitch-T (DINOv3). Overall, both qualitative and quantitative results demonstrate that VFMStitch provides a more robust and accurate solution for large-motion 3DUS stitching.

Effectiveness of intensity-based refinement: We further evaluate the effectiveness of the intensity-based refinement applied after VFMStitch. As reported in Table 1, ANTs-based refinement consistently improves pixel-wise alignment as measured by NCC.

Fig. 4 presents a qualitative comparison before and after refinement. While VFMStitch alone already achieves good global alignment, the subsequent intensity-based refinement further enhances local voxel correspondence. Importantly, the failure of ANTs when applied independently highlights that this improvement critically depends on the strong global initialization provided by VFMStitch. Purely intensity-based registration struggles under large inter-volume motion, whereas its effectiveness is recovered when guided by a robust VFMStitch-based initial alignment. These results confirm that intensity-based refinement serves as a complementary post-processing step, improving fine-scale alignment without altering the robustness or training-free nature of the proposed framework.

Effectiveness of feature fusion: Table 2 compares the performance of geometric descriptors alone to fusing geometric and semantic features. We observe that VFMStitch equipped with fused descriptors (red) consistently outperforms both the semantic-only configuration (DINOv3, green) and purely geometric baselines (gray) with the exception of

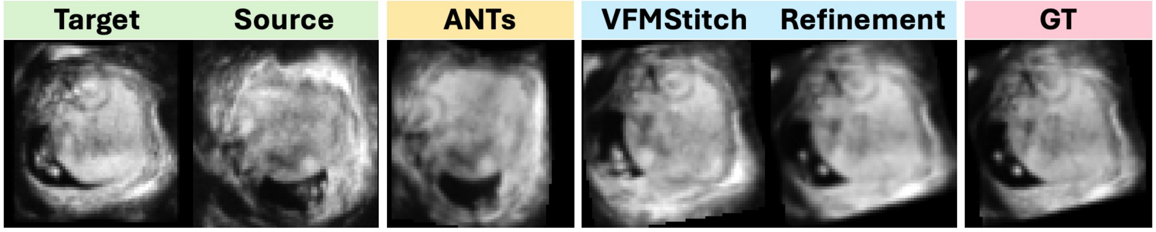

Figure 4: Comparison of registration performance before and after intensity-based refinement using ANTs (Rigid) for VFMStitch (DINOv3+FPFH, T). VFMStitch provides accurate global alignment, which can be further refined by ANTs to improve local registration accuracy. In contrast, ANTs (Rigid) alone fails to achieve satisfactory alignment, highlighting the importance of VFMStitch-based initialization.

Table 2: Comparison of geometric descriptor performance before (gray) and after fusion with DINOv3 semantic features (red), with DINOv3-only results shown in green. RANSAC and ANTs are used for matching and refinement, respectively. **Best** and second best results are highlighted. Geometric–semantic fusion consistently outperforms both geometric-only and semantic-only representations, demonstrating the effectiveness of the proposed fusion strategy.

Methods	NCC ↑	MSE(×10) ↓	PSNR ↑	SSIM ↑
FPFH (R, F)	0.889 ± 0.125	0.105 ± 0.098	23.256 ± 6.602	0.654 ± 0.262
Point-BERT (R, F)	0.871 ± 0.123	0.131 ± 0.110	21.910 ± 6.526	0.605 ± 0.251
Point-DAE (R, F)	0.868 ± 0.122	0.131 ± 0.106	21.397 ± 5.693	0.597 ± 0.244
Point-MAE (R, F)	0.866 ± 0.120	0.137 ± 0.110	21.310 ± 5.805	0.592 ± 0.245
VFMStitch(DINOv3, R, F)	$\underline{0.910 \pm 0.114}$	0.090 ± 0.097	24.064 ± 6.542	0.690 ± 0.242
VFMStitch(DINOv3+FPFH, R, F)	0.906 ± 0.153	0.087 ± 0.113	24.346 ± 6.224	$\underline{0.709 \pm 0.243}$
VFMStitch(DINOv3+Point-BERT, R, F)	0.910 ± 0.126	0.087 ± 0.099	$\mathbf{24.779 \pm 6.932}$	0.708 ± 0.248
VFMStitch(DINOv3+Point-DAE, R, F)	0.908 ± 0.125	$\underline{0.085 \pm 0.092}$	24.474 ± 6.581	0.705 ± 0.243
VFMStitch(DINOv3+Point-MAE, R, F)	$\mathbf{0.916 \pm 0.116}$	$\mathbf{0.084 \pm 0.095}$	$\underline{24.656 \pm 6.586}$	$\mathbf{0.712 \pm 0.241}$

a slight decrease in NCC when fusing DINOv3 with FPFH or Point-DAE. Importantly, this slight reduction does not affect structural similarity or overall registration accuracy, as reflected by the consistent gains in SSIM, PSNR, and MSE. These results suggest that combining low-level geometric structure with high-level semantic context yields more robust and discriminative point correspondences, ultimately improving registration performance in challenging large-motion 3DUS stitching scenarios.

Effectiveness of DINOv3-based descriptor: Fig. 5 presents a quantitative comparison between DINOv3-based semantic descriptors and baseline geometric descriptors using different PCD matchers. Across both matching strategies, the DINOv3-based purely semantic descriptor consistently yields superior performance compared to purely geometric descriptors. This trend is consistent across all evaluated metrics, indicating that DINOv3-derived features provide more discriminative and robust representations for establishing reliable

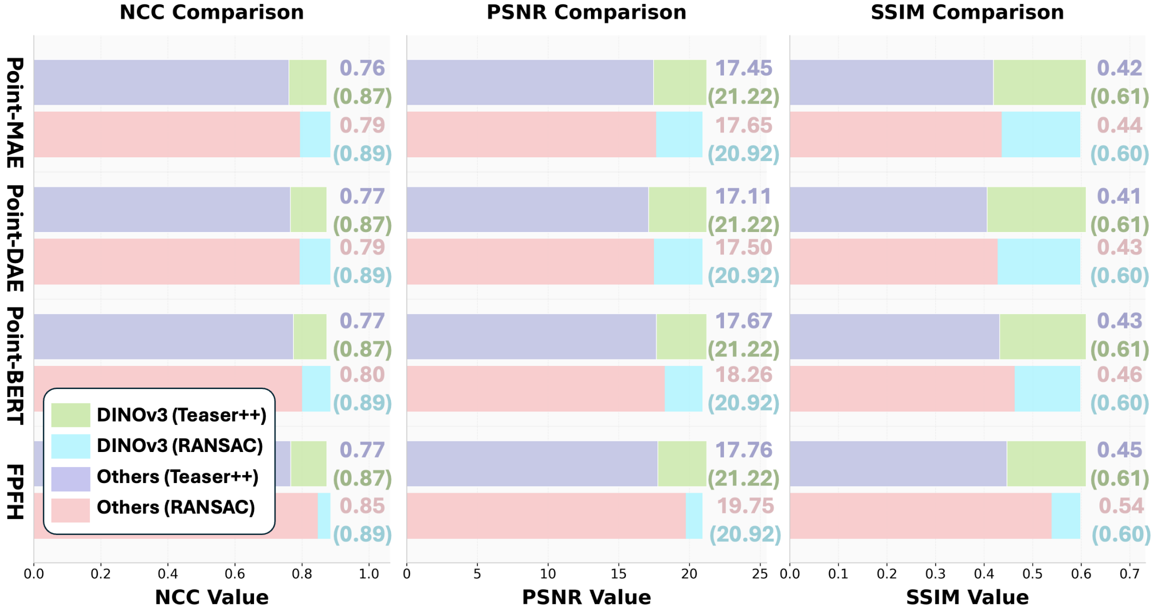

Figure 5: Comparison between DINOv3 and the geometric descriptors using RANSAC and TEASER++ as PCD matchers. Across all evaluated metrics, DINOv3 consistently achieves superior performance, demonstrating its effectiveness as a semantic descriptor for robust point-cloud registration.

correspondences under large rigid motion. We attribute this advantage to the semantic-rich and spatially coherent representations learned by DINOv3 through large-scale self-supervised pretraining, which are better suited to capturing high-level anatomical context than descriptors optimized solely for local geometric patterns. These results demonstrate that DINOv3-derived semantic features can substantially enhance registration robustness in challenging 3DUS stitching scenarios.

Discussion and conclusion: In this work, we presented **VFMStitch**, the first training-free framework that systematically integrates vision foundation model–derived semantic features with PCD geometric registration for large-motion 3DUS stitching. By explicitly reformulating 3DUS stitching as a robust rigid registration problem in the PCD domain, VFMStitch overcomes the fundamental limitations of the current VFM-based approaches under large translations and rotations. Extensive quantitative and qualitative evaluations demonstrate that VFMStitch consistently outperforms SOTA conventional, learning-based, and VFM-based baselines, highlighting its effectiveness for dealing with challenging ultrasound data. Moreover, we show that VFMStitch provides a robust initialization for subsequent intensity-based refinement, enabling improved pixel-level alignment without sacrificing robustness. Beyond 3DUS stitching, our results suggest a broader potential for leveraging foundation model features in rigid and rigid medical image registration tasks, particularly in scenarios where large motion and low image quality challenge traditional intensity-driven methods. We believe this work opens new directions for training-free, foundation-model-guided geometric registration in medical imaging.

Acknowledgments

This work is supported, in part, by NIH R01-HD109739, R01-HL156034, T32-EB021937, and the Vanderbilt Advanced Computing Center for Research and Education.

References

Jakob Ambsdorf, Asbjørn Munk, Sebastian Llambias, Anders N Christensen, Kamil Mikolaj, Randall Balestriero, Martin G Tolsgaard, Aasa Feragen, and Mads Nielsen. General methods make great domain-specific foundation models: A case-study on fetal ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 271–281. Springer, 2025.

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

J Banerjee, C Klink, ED Peters, WJ Niessen, A Moelker, and T van Walsum. Fast and robust 3D ultrasound registration–block and game theoretic matching. *MedIA*, 20(1): 173–183, 2015.

S Bano and D Stoyanov. Image mosaicking. *MedIA*, pages 387–411, 2024.

Paul J Besl and Neil D McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

Chengqian Che, Tejas Sudharshan Mathai, and John Galeotti. Ultrasound registration: A review. *Methods*, 115:128–143, 2017.

Robert R Entrekin, Bruce A Porter, Henrik H Sillesen, Anthony D Wong, Peter L Cooperberg, and Cathy H Fix. Real-time spatial compound imaging: application to breast, vascular, and musculoskeletal ultrasound. In *Seminars in Ultrasound, CT and MRI*, volume 22-1, pages 50–64. Elsevier, 2001.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

A Gomez, V Zimmer, N Toussaint, R Wright, JR Clough, B Khanal, MPM van Poppel, E Skelton, J Matthews, and JA Schnabel. Image reconstruction in a manifold of image patches: Application to whole-fetus ultrasound imaging. In *Machine learning for medical image reconstruction*, pages 226–235. Springer, 2019.

Alberto Gomez, Kanwal Bhatia, Sarjana Tharin, James Housden, Nicolas Toussaint, and Julia A Schnabel. Fast registration of 3d fetal ultrasound images using learned corresponding salient points. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 33–41. Springer, 2017.

Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson, 3rd edition, 2008.

Hanxue Gu, Yaqian Chen, Nicholas Konz, Qihang Li, and Maciej A Mazurowski. Are vision foundation models ready for out-of-the-box medical image registration? In *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care*, pages 101–112. Springer, 2025.

Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Manchester, UK, 1988.

Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical image analysis*, 16(7):1423–1435, 2012.

Zhongliang Jiang, Chenyang Li, Xuesong Lil, and Nassir Navab. Thoracic cartilage ultrasound-ct registration using dense skeleton graph. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6586–6592. IEEE, 2023.

Zhongliang Jiang, Yunfeng Kang, Yuan Bi, Xuesong Li, Chenyang Li, and Nassir Navab. Class-aware cartilage segmentation for autonomous us-ct registration in robotic intercostal ultrasound imaging. *IEEE Transactions on Automation Science and Engineering*, 22:4818–4830, 2024.

Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2): 358–367, 1988.

J P Lewis. Fast normalized cross-correlation. In *Vision Interface*, volume 10, pages 120–123, 1995.

Yuheng Li, Yizhou Wu, Yuxiang Lai, Mingzhe Hu, and Xiaofeng Yang. Meddinov3: How to adapt vision foundation models for medical image segmentation? *arXiv preprint arXiv:2509.02379*, 2025.

David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980.

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

Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 604–621. Springer, 2022.

Blaine Rister, Mark A Horowitz, and Daniel L Rubin. Volumetric image registration from invariant keypoints. *IEEE Transactions on Image Processing*, 26(10):4900–4910, 2017.

ME Roy-Lacroix, F Moretti, ZM Ferraro, L Brosseau, J Clancy, and K Fung-Kee-Fung. A comparison of standard two-dimensional ultrasound to three-dimensional volume sonography for routine second-trimester fetal imaging. *Journal of Perinatology*, 37(4):380–386, 2017.

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009.

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

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jiyong Tan, Hui Qin, Xinxing Chen, Jiawang Li, Yuanwei Li, Bing Li, Yuquan Leng, and Chenglong Fu. Point cloud segmentation of breast ultrasound regions to be scanned by fusing 2d image instance segmentation and keypoint detection. In *2023 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 669–674. IEEE, 2023.

Jihang Wang, Vikas Shivaprabhu, John Galeotti, Samantha Horvath, Vijay Gorantla, and George Stetten. Towards video guidance for ultrasound, using a prior high-resolution 3d surface map of the external anatomy. In *Augmented Environments for Computer-Assisted Interventions: 9th International Workshop, AE-CAI 2014, Held in Conjunction with MICCAI 2014, Boston, MA, USA, September 14, 2014. Proceedings 9*, pages 51–59. Springer, 2014.

Shansong Wang, Mojtaba Safari, Mingzhe Hu, Qiang Li, Chih-Wei Chang, Richard LJ Qiu, and Xiaofeng Yang. Dinov3 with test-time training for medical image registration. *arXiv preprint arXiv:2508.14809*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

R Wright, N Toussaint, A Gomez, V Zimmer, B Khanal, J Matthew, E Skelton, B Kainz, D Rueckert, JV Hajnal, and JA Schnabel. Complete fetal head compounding from multi-view 3DUS. *MICCAI*, pages 384–392, 2019.

R Wright, A Gomez, V A Zimmer, N Toussaint, B Khanal, J Matthew, E Skelton, B Kainz, D Rueckert, J V Hajnal, and JA Schnabel. Fast fetal head compounding from multi-view 3DUS. *MedIA*, 2023.

H. Yang, J. Shi, and L. Carlone. TEASER: Fast and Certifiable Point Cloud Registration. *IEEE Trans. Robotics*, 2020.

X Yao, H Liu, D Hu, D Lu, A Lou, H Li, R Deng, G Arenas, B Oguz, N Schwartz, B Byram, and I Oguz. FNPC-SAM: uncertainty-guided false negative/positive control for SAM on noisy medical images. In *SPIE Medical Imaging*, volume 12926, 2024.

Xing Yao, Runxuan Yu, Nick DiSanto, Ehsan Khodapanah Aghdam, Kanyifeechukwu Jane Oguine, Daiwei Lu, Ange Lou, Jiacheng Wang, Dewei Hu, Gabriel A Arenas, et al. Lotus: Latent outpainting diffusion model for three-dimensional ultrasound stitching. In *Medical Imaging with Deep Learning*, 2025a.

Xing Yao, Runxuan Yu, Dewei Hu, Hao Yang, Ange Lou, Jiacheng Wang, Daiwei Lu, Gabriel Arenas, Baris Oguz, Alison Pouch, et al. Synstitch: A self-supervised learning network for ultrasound image stitching using synthetic training pairs and indirect supervision. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025b.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yabin Zhang, Jiehong Lin, Ruihuang Li, Kui Jia, and Lei Zhang. Point-dae: Denoising autoencoders for self-supervised point cloud learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018.