

SAMIR, an efficient registration framework via robust feature learning from SAM

Yue He^{1*}, Min Liu^{1*}, Qinghao Liu¹, Jiazheng Wang¹, Yaonan Wang¹, Hang Zhang¹, Xiang Chen^{1 †}

¹College of Electrical and Information Engineering, Hunan University, Changsha, Hunan, China

Abstract

Image registration is a fundamental task in medical image analysis. Deformations are often closely related to the morphological characteristics of tissues, making accurate feature extraction crucial. Recent weakly supervised methods improve registration by incorporating anatomical priors such as segmentation masks or landmarks, either as inputs or in the loss function. However, such weak labels are often not readily available, limiting their practical use. Motivated by the strong representation learning ability of visual foundation models, this paper introduces SAMIR, an efficient medical image registration framework that utilizes the Segment Anything Model (SAM) to enhance feature extraction. SAM is pretrained on large-scale natural image datasets and can learn robust, general-purpose visual representations. Rather than using raw input images, we design a task-specific adaptation pipeline using SAM's image encoder to extract structure-aware feature embeddings, enabling more accurate modeling of anatomical consistency and deformation patterns. We further design a lightweight 3D head to refine features within the embedding space, adapting to local deformations in medical images. Additionally, we introduce a Hierarchical Feature Consistency Loss to guide coarse-to-fine feature matching and improve anatomical alignment. Extensive experiments demonstrate that SAMIR significantly outperforms state-of-the-art methods on benchmark datasets for both intra-subject cardiac image registration and inter-subject abdomen CT image registration, achieving performance improvements of 2.68% on ACDC and 6.44% on the abdomen dataset. The source code will be publicly available on GitHub following the acceptance of this paper.

Introduction

Pair-wise image registration is essential in medical image processing, aligning moving and fixed images from different subjects, modalities, time points, or perspectives to enhance diagnosis, surgical planning, and motion analysis (Chen et al. 2021a).

Traditional registration methods rely on iterative optimization to minimize the distance like normalized cross-correlation (NCC) or mean square error (MSE), between warped moving and fixed images, which is time-consuming.

*These authors contributed equally.

†Corresponding authors: xiangc@hnu.edu.cn

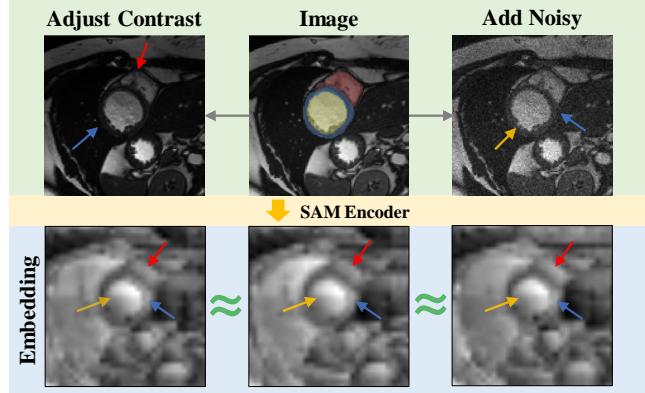


Figure 1: Feature extraction from SAM encoder for different scenarios of images. It can be observed that the embeddings of the SAM image encoder show strong robustness, accurately capturing object structures even with contrast changes and noise interference.

Deep learning-based registration has significantly accelerated this process, enabling registration through a single forward inference after training. However, these methods typically employ conventional loss functions designed to minimize raw image contrast discrepancies. Crucially, they fail to take into account inherent variations in image quality (e.g., contrast heterogeneity and noise patterns) that arise from differences in imaging equipment, acquisition protocols, and operator techniques, even for repeated scans of the same subject.

Current deep learning methods enhance feature learning in registration networks through various architectures, either explicitly or implicitly. Early studies used convolutional layers for feature extraction and integration (Balakrishnan et al. 2019; Dalca et al. 2019a), and later incorporated transformers to capture long-range correlations (Chen et al. 2022a). Following research have also explored the large kernel convolution (Jia et al. 2022a) and multi-layer perceptron (Meng et al. 2024) to further enhance the feature learning in the registration network. Some approaches also proposed to learn explicit representations from the input images and then fed them into the registration block (Lee et al. 2019). For example, Lee et al. (Lee et al. 2019) proposed an

image-and-spatial transformer network to leverage the structure of information to learn new image representations that are optimized for the downstream registration task. Previous research has demonstrated that incorporating anatomical structure priors like segmentation masks or landmarks in network training can lead to significantly better registration performance (Balakrishnan et al. 2019; Chen et al. 2021b; Young et al. 2022). These approaches, also known as the weakly-supervised registration methods, either utilize segmentation/landmarks as a loss function in the network training (Balakrishnan et al. 2019), or directly incorporate them as the input of the network (Chen et al. 2021b). However, weak labels such as segmentation masks are not always available, limiting their practical applications.

Recently, the Segment Anything Model (SAM)(Kirillov et al. 2023) has significantly revolutionized the field of image segmentation. One of its key strength lies in generalization capabilities, enabled by training on a diverse dataset with extreme lighting, complex occlusion, and boundary uncertainties. This allows SAM to achieve impressive zero-shot performance across tasks, often matching or surpassing fully supervised results. In addition, the feature embeddings learning from the SAM encoder exhibit strong robustness to either contrast variations or noise, as shown in Figure 1. Fine-tuning SAM on medical datasets has further enhanced its segmentation capabilities(Cheng et al. 2023; Ma et al. 2024; Yue et al. 2024; Wenhui et al. 2025). While SAM has achieved success in numerous downstream tasks, its application in the field of image registration remains limited.

To the best of our knowledge, there are currently only two works that apply SAM to medical image registration. One approach is SAMReg (Huang et al. 2024), which leveraged SAM for multi-class segmentation of image pairs instead of directly predicting a global deformation field, aiming to achieve accurate local registration. Another is SAM-Assisted Registration (Xu et al. 2025), which utilized text prompts to guide SAM in generating segmentation masks and then employed them to assist both training and inference, ensuring anatomical consistency. While they do not directly utilize SAM in the deformation fields prediction, their efforts demonstrate the promising potential for further exploration in this field.

In this paper, we propose **SAMIR** (Segment Anything Model for Image Registration), a novel medical image registration framework that leverages SAM to enhance feature learning, aiming to achieve more robust and accurate image alignment. SAMIR introduces a registration-specific adaptation pipeline for SAM, utilizing its powerful pretrained image encoder and complementing it with a lightweight 3D head to extract accurate, fine-grained, and robust image features. Furthermore, a hybrid loss function is introduced to align the moving and fixed images across multi-scale spaces, addressing the challenges of robust feature learning and large deformation modeling.

In summary, the contributions of this paper can be summarized as follows:

- We proposed a novel feature-driven registration framework, SAMIR, leveraging the structure-aware properties of foundation models to achieve robust and accurate

medical image registration.

- A novel feature-level loss is designed to further enhance the structure alignment and robustness on registration.
- Our SAMIR achieves state-of-the-art (SOTA) performance on multiple registration tasks, including the ACDC dataset and abdomen dataset, with Dice scores improved by 2.68% and 6.44%, respectively.

Related Works

Deformable Medical Image Registration

Early image registration networks typically employ a single-step approach for deformation field estimation. For example, VoxelMorph(Balakrishnan et al. 2019) adopts an end-to-end U-Net architecture to predict the deformation fields. Such types of methods can handle small local deformation well, while they often struggle with large deformations. Following research like TransMorph(Chen et al. 2022a) and LKUNet(Jia et al. 2022a) proposed to replace convolutional modules with Transformer components and large kernel convolution, to enlarge the perceptual fields. Despite improved accuracy, large deformation registration remains challenging. Recent work has shifted toward more sophisticated architectures to improve accuracy and robustness. The main approaches fall into two categories. The first category is adapting cascaded structures. For instance, VTN (Zhao et al. 2019), which utilized recursive cascaded networks, splitting the registration task into a multi-step cascade small deformation. VR-Net (Jia et al. 2022b) splits image registration into closed-form and denoising subproblems, models them with specialized layers, and cascades them for fast, accurate, data-efficient registration. The second category is based on coarse-to-fine pyramid strategies. For instance, LapIRN(Mok and Chung 2020a) introduced a deep Laplacian pyramid network to progressively refine deformations from low to high resolution. RPD(Wang, Ni, and Wang 2024) integrated recursion into the pyramid framework to better handle large displacements, while CorrMLP(Meng et al. 2024) presented the first MLP-based coarse-to-fine registration model. These strategies enhance multi-scale modeling and substantially enhance complex registration performance.

Visual Foundation Models

Foundation models have recently seen rapid development in computer vision, achieving strong generalization capabilities and good adaptability to a wide range of downstream tasks through pre-training on large-scale and diverse datasets. Segment Anything (SAM)(Kirillov et al. 2023) demonstrated impressive zero-shot performance across various image segmentation tasks via pre-training on over one billion masks. SegGPT(Wang et al. 2023) unified image segmentation into a general visual perception task by transforming diverse segmentation tasks into identically formatted images and incorporating a random color mapping mechanism for contextual coloring, thereby establishing a context-based learning framework for universal segmentation. DINO(Caron et al. 2021) addressed unsupervised visual representation learning via self-distillation

and contrastive learning with a student-teacher architecture and pseudo-labeling, enabling transferable feature learning from large-scale unlabeled data and achieving strong performance. DINOv2(Oquab et al. 2023) combined DINO and iBOT(Zhou et al. 2022) with KoLeo regularization and Sinkhorn-Knopp centering to learn robust visual features without supervision, achieving strong performance on classification, segmentation, and retrieval tasks.

While foundation models have shown considerable success in computer vision, their application to medical image registration remains challenging due to: (1)the lack of reliable ground-truth deformation fields for training and validation, and (2) the scarcity of publicly available, well-paired medical image datasets, further complicated by their large volumetric dimensions. Notably, existing research (Dong et al. 2017; Chen et al. 2022b) has demonstrated that registration and segmentation share core feature-matching principles, enabling mutual reinforcement. Capitalizing on this theoretical connection, we propose SAMIR, a novel framework that leverages SAM’s structure-aware properties to guide deformation field estimation in registration tasks.

Method

Given paired moving image I_m and fixed image I_f defined over the spatial domain $\Omega \subseteq R^3$, image registration aims to estimate a deformation field $\phi(\circ) : R^3 \rightarrow R^3$ that optimally aligns I_m to I_f , such that $I_m \circ \varphi \approx I_f$. The deformation field is typically formulated as,

$$\phi(\mathbf{x}) = \mathbf{x} + u(\mathbf{x}), \quad (1)$$

where, \mathbf{x} represents voxel coordinates, and $u(\mathbf{x})$ denotes the displacement field.

As illustrated in Figure 2, the SAMIR framework comprises two principal components: (1) A Structure-Aware Feature Embedding (SAFE) module that extracts robust structural intermediate representations from both I_m and I_f using the encoder from a visual foundation model, subsequently performing feature integration and medical domain adaptation via a 3D convolutional head; (2) A Pyramid Deformation Field Prediction (PDFP) module employing a coarse-to-fine progressive optimization strategy to generate high-precision displacement fields adaptable to various deformation magnitudes. Notably, the foundation model encoder remains frozen throughout training, substantially reducing trainable parameters while conserving computational resources and training time.

Structure-Aware Feature Embedding

In this paper, we utilize SAM as the foundation model of our SAMIR for feature encoding. The SAM image encoder is a Vision Transformer (ViT) pretrained via Masked autoencoders (Kaiming et al. 2021) on a large and diverse dataset, endowing it with robust structure-aware properties and powerful zero-shot transfer performance.

Dimensional Adaptation of Foundation Models. To adapt the fundamental vision model for medical image registration, we address several critical discrepancies between standard natural images and medical imaging data. (1)Current foundation models such as SAM are predominantly

Algorithm 1: registration-specific Adaptation Pipeline

Input: Image $I \in R^{H \times W \times D}$
Parameter: Padding size (H', W') , Target size k
Output: Feature embeddings F_{final}

```

1: Preprocess Input Slices
2: for  $i = 1$  to  $D$  do
3:    $I^i \leftarrow \text{Slice}(I, i)$  {Extract  $i$ -th slice}
4:    $\hat{I}^i \leftarrow \text{Pad}(I^i, (H', W'))$ 
5:    $I_{\text{input}}^i \leftarrow \text{Up}(\hat{I}^i, (k, k))$ 
6: end for
7: Extract and Stack Features
8: for  $i = 1$  to  $D$  do
9:    $f^i \leftarrow \text{SAM\_encoder}(I_{\text{input}}^i)$ 
10: end for
11:  $F_{\text{ori}} \leftarrow \text{Stack}(f^1, \dots, f^D)$ 
12: Restore Spatial Dimensions
13:  $F_{\text{up}} \leftarrow \text{Up}(F_{\text{ori}}, (B, C, H', W', D))$ 
14:  $F_{\text{general}} \leftarrow \text{Crop}(F_{\text{up}}, (B, C, H, W, D))$ 
15: Feature enhancement
16:  $F_{\text{final}} \leftarrow \text{3D\_head}(F_{\text{general}})$ 
17: return  $F_{\text{final}}$ 

```

designed for 2D natural images, whereas medical imaging largely operates in 3D space. (2)While SAM requires square input dimensions ($H = W$), medical images frequently exhibit arbitrary aspect ratios. (3) A significant resolution gap exists between SAM’s training data (shorter side resized to 1500 pixels) and medical images (generally lower resolutions such as 128×128). To bridge these gaps, we implement a multi-stage adaptation strategy. For 3D compatibility, we follow (Chen et al. 2024a), processing volumetric medical data $[B, H, W, D]$ as sequential 2D slices $[BD, H, W]$, extracting features slice-wise $F_{\text{slice}} = \{f^1, f^2, \dots, f^D\}, f^i \in R^{C \times H \times W}$, and then reorganizing them into coherent 3D embeddings F_{ori} . This approach effectively leverages the 2D backbone while preserving volumetric information. For handling varying aspect ratios, we employ strategic padding from $[H, W, D]$ to $[H', W', D]$ to conform to SAM’s square input requirements, recording padding parameters to enable precise restoration to original dimensions during post-processing. Resolution mismatches are mitigated through intelligent upsampling of medical images from $[H', W', D]$ to $[k, k, D]$ to match the encoder’s expected input dimensions prior to feature extraction.

In addition, the feature embeddings of the SAM encoder are at 1/16th the spatial resolution of the original input, which leads to loss of spatial information and inaccurate prediction of deformation fields. To tackle this issue, we apply a two-stage upsample operation (the inputs are upsampled to $(k/H') \times$ prior to the encoder, and the features are upsampled $(16H/k) \times$) to upsample the obtained feature embeddings to match the input size, resulting in a dimension of $[256, H, W, D]$.

Domain Adaptation of Foundation Models. After extracting structure-aware feature representation from both fixed and moving images, we develop a lightweight 3D con-

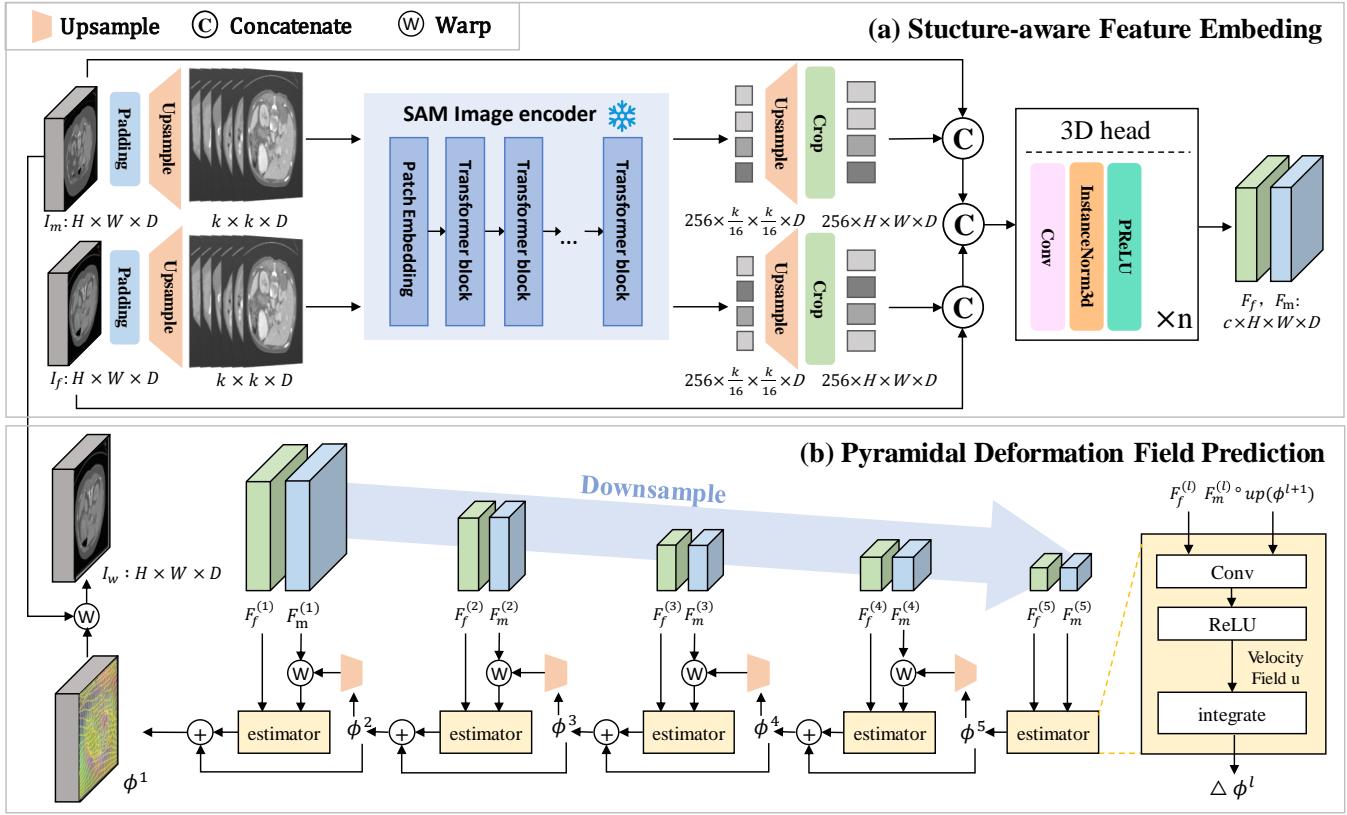


Figure 2: Overview of SAMIR Network Architecture. The SAMIR network leverages the SAM pre-trained image encoder to extract stucture-aware feature embeddings and employs a pyramidal deformation field prediction block to achieve coarse-to-fine progressive displacement field optimization, ultimately accomplishing accurate cross-modality medical image registration.

volutional module to generate enhanced feature representations $[C, H, W, D]$. Considering that foundation model encoders produce high-dimensional features (e.g., SAM’s 256-channel output) that increase computational costs, and that conventional 2D slice-wise processing compromises volumetric consistency, this module uses 3D convolutions to jointly reduce computational complexity through efficient feature fusion and improve spatial continuity by enhancing inter-slice correlations, enabling better capture of medical image features for improved registration performance.

The overall workflow can be formulated in Algorithm 1. With these simple techniques and domain adaptation, the visual foundation model achieves registration-specific adaptation without introducing a large number of additional parameters.

Pyramidal Deformation Field Prediction

To handle both small and large deformations, coarse-to-fine registration based on pyramid features is an efficient scheme. Following previous research (Chen et al. 2024b), the input feature embeddings are first downsampled hierarchically to generate multi-scale feature maps. Subsequently, the displacement field is progressively optimized at each level, from coarse to fine. In each layer, we take the moving features and fixed features as inputs and utilize a three-layer

convolution block to predict the corresponding velocity field \mathbf{u} . After integration (Dalca et al. 2019b), the deformation field ϕ_0^l can be obtained. If there is a lower layer, the resultant deformation fields ϕ^{l+1} would be upsampled to the same resolution as the current layer and composed with ϕ_0^l for progressive refinement, formulated as,

$$\begin{aligned}\tilde{\phi}^{l+1} &= \text{up}(\phi^{l+1}), \\ \phi^l &= \tilde{\phi}^{l+1} \circ \Delta\phi_0^l,\end{aligned}\quad (2)$$

where $\text{up}(\cdot)$ denotes $2\times$ trilinear upsampling and scaling, $\exp(\cdot)$ refers to the scaling and squaring function applied to the displacement field, and \circ denotes the warping function. Note that, on the bottom layer, the deformation field ϕ^l is exactly ϕ_0^l . This design captures global deformation patterns at coarse levels and optimizes local details at fine levels, thereby achieving globally precise registration.

Hierarchical Loss Function

In medical image registration, single-scale similarity metrics often struggle to achieve a balance between global alignment accuracy and the preservation of local anatomical details. Previous approaches typically compute intensity differences at the image level within a multi-scale pyramid framework, overlooking the high-level semantic consistency embedded

in deep feature spaces, and critically depend on highly consistent image conditions. The potential differences in scanning conditions between moving and fixed images may result in local misalignments when using intensity-only loss functions, particularly in cases involving complex anatomical structures or images affected by noise artifacts.

To address this issue and enhance the overall alignment consistency, we introduce a hierarchical feature consistency loss, denoted as L_{HFC} . Specifically, the feature embeddings extracted by the SAM encoder from both the fixed and moving images are downsampled to generate multi-level feature representations, denoted as F_{fix} and F_{mov} , respectively. At each resolution level l , the moving image features are first spatially transformed using the corresponding deformation field Φ^l , and then the similarity discrepancy is computed with respect to the fixed image features at the same scale. The formulation of the hierarchical feature consistency loss is given as follows:

$$L_{HFC} = \sum_{l=1}^n \frac{1}{2^{l-1}} \|F_{mov}^l \circ \phi^l - F_{fix}^l\|. \quad (3)$$

Overall Losses

Following (Chen et al. 2024c; Cheng et al. 2025), SAMIR uses a loss with a dissimilarity term and a regularization term. The dissimilarity loss combines image-level normalized cross-correlation (NCC) L_{NCC} and feature-level hierarchical feature consistency loss L_{HFC} . For the weakly supervised version, where segmentation masks are used during network training, we also compute regional dissimilarity using the Dice loss L_{Dice} , as proposed in (Chen et al. 2021b).

$$L_{sim} = \lambda_0 L_{NCC} + \lambda_1 L_{HFC} + \lambda_2 L_{Dice}, \quad (4)$$

Similar to previous works (Dalca et al. 2019b), we adopt a diffusion-based smoothness regularization term $L_{smooth} = \sum_{x \in \Omega} \|\nabla \varphi(x)\|^2$ to enforce spatial smoothness of the deformation field.

The complete loss function is formulated as,

$$L_{total} = L_{sim} + \lambda_3 L_{smooth}. \quad (5)$$

where, λ_0 , λ_1 , λ_2 and λ_3 are hyperparameters to balance different losses.

Experiment

Experiment Setting

Datasets and Implementation Details. We evaluate our SAMIR on two datasets: the ACDC cardiac MRI dataset (Bernard et al. 2018) and Abdomen CT dataset (Xu et al. 2016). The ACDC dataset focuses on intra-subject motion tracking between end-diastole (ED) and end-systole (ES) phases, comprising 80 training, 20 validation, and 50 test cases. Bidirectional registration (ED-to-ES and ES-to-ED) yields 160, 40, and 100 image pairs for training, validation, and testing, respectively. All images were preprocessed to $128 \times 128 \times 16$ (resolution $1.8 \times 1.8 \times 10 mm^3$). The Abdomen CT dataset, from the Learn2Reg challenge, addresses

inter-subject registration with large deformations across abdominal organs (*e.g.*, liver, kidneys, spleen, pancreas). It includes 30 CT scans: 20 training, 3 validation, and 7 test cases. Pairwise combinations created 380 training, 6 validation, and 42 test pairs, processed to $192 \times 160 \times 256$. SAM features were precomputed to avoid redundant extraction. The experiments utilized the Adam optimizer with a learning rate of $1e^{-4}$. All experiments were implemented in PyTorch and executed on a single NVIDIA RTX A6000 GPU. The hyper-parameters $\lambda_0 - \lambda_3$ are all set to 1.

Comparison Methods and Evaluation Metrics. We compared our method with SOTA deformable image registration approaches, including VoxelMorph(Balakrishnan et al. 2018), TransMorph(Chen et al. 2022a), LKUNet(Jia et al. 2022a), Fourier-Net(Jia et al. 2023), CorrMLP(Meng et al. 2024), MemWarp (Zhang et al. 2024), and RDP(Wang, Ni, and Wang 2024), where the latter three using feature pyramid architectures. For both ACDC and abdominal datasets, we used publicly available codes and fine-tuned each model for optimal performance. For the abdominal dataset, we additionally evaluated LapIRN (Mok and Chung 2020b)(incompatible with short-axis data), ConvexAdam(Siebert, Hansen, and Heinrich 2021), and SAMConvex(Li et al. 2023) (effective for large deformations). Following (Dalca et al. 2019b; Chen et al. 2022a; Zhang et al. 2024; Chen et al. 2024c), we employed the Dice Similarity Coefficient (Dice) and the 95% Hausdorff Distance (HD95) to assess anatomical alignment, and standard deviation of the Jacobian determinant logarithm (SDlogJ) to evaluate the smoothness of deformation fields. Computational efficiency was measured by multi-adds (MAs), parameter size (PS), and average inference time.

Results and Analysis

Intra-subject Registration on ACDC. As shown in Table 1, the pyramid-based methods (CorrMLP, MemWarp, RDP) consistently outperform single-layer deformation approaches, including traditional convolution (VoxelMorph), transformer-based (TransMorph), large-kernel convolution (LKU-Net), and Fourier-Net architectures. This highlights the effectiveness of pyramid-based methods in capturing both small and large deformations. With robust structure-aware feature learning from SAM, SAMIR significantly outperforms those pyramid-based registration methods ($p < 0.05$), with a 2.68% improvement over the sub-optimal method RDP. Note that SAM features can be precomputed, requiring only 0.94 seconds per ACDC sample for extraction, and thus impose no additional computational burden during registration.

Inter-subject Registration on Abdomen Dataset. We demonstrated SAMIR’s effectiveness in large deformation tasks on abdominal CT, as shown in Tabla 2 and Figure 3. Similar to the results on ACDC, the pyramid-based registration methods significantly outperform VoxelMorph, TransMorph, and LKUNet, as the deformation across different subjects is larger than intra-subject registration. Our SAMIR achieves significantly better registration performance than them, with a 6.44% improvement on the average Dice

| Model | Dice (%) ↑ | HD95 (mm) ↓ | SDlogJ ↓ | MAs (G) ↓ | PS (MB) ↓ | Time ↓ |
|---------------------------------------|--------------|-------------|----------|-----------|-----------|--------|
| Initial | 58.14 | 11.95 | - | - | - | - |
| VoxelMorph (Balakrishnan et al. 2018) | 75.26 | 9.33 | 0.044 | 19.5 | 0.32 | 0.18 |
| TransMorph (Chen et al. 2022a) | 74.97 | 9.44 | 0.045 | 50.20 | 46.69 | 0.26 |
| LKU-Net (Jia et al. 2022a) | 76.53 | 9.13 | 0.049 | 160.50 | 33.35 | 0.22 |
| Fourier-Net (Jia et al. 2023) | 76.61 | 9.25 | 0.047 | 86.07 | 17.43 | 0.27 |
| CorrMLP (Meng et al. 2024) | 77.31 | 9.00 | 0.056 | 47.59 | 4.19 | 0.28 |
| MemWarp (Zhang et al. 2024) | 76.74 | 9.67 | 0.108 | 1270.00 | 47.78 | 0.58 |
| RDP (Wang, Ni, and Wang 2024) | 78.06 | 9.02 | 0.076 | 154.00 | 8.92 | 0.36 |
| SAMIR-vith (Ours) | 80.74 | 8.22 | 0.048 | 230.34 | 7.15 | 0.32 |

Table 1: Quantitative comparison on the cardiac ACDC dataset. Statistically significant improvements in registration accuracy are highlighted in bold. Symbols indicate direction: ↑ for higher is better, ↓ for lower is better. “Initial” refers to baseline results before registration.

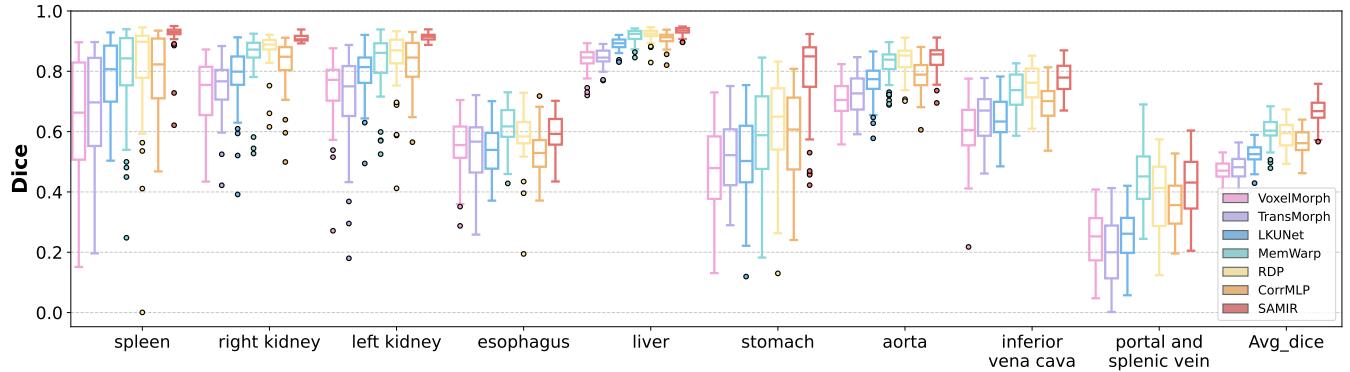


Figure 3: Boxplot on the abdomen CT dataset. Our SAMIR significantly outperforms the rest approaches on all Dice scores.

| Model | Dice (%) ↑ | HD95 (mm) ↓ | SDlogJ ↓ |
|--------------------------|--------------|--------------|----------|
| Initial | 30.68 | 29.77 | - |
| VoxelMorph | 47.05 | 23.08 | 0.13 |
| TransMorph | 47.94 | 21.53 | 0.13 |
| LKUNet | 52.78 | 20.56 | 0.98 |
| LapIRN | 54.55 | 20.52 | 1.73 |
| CorrMLP | 56.11 | 19.52 | 0.16 |
| RDP | 58.77 | 20.07 | 0.22 |
| MemWarp | 60.24 | 19.84 | 0.53 |
| FourierNet | 42.80 | 22.95 | 0.13 |
| ConvexAdam | 51.10 | 23.14 | 0.11 |
| SAMConvex | 53.65 | 18.66 | 0.12 |
| SAMIR-vith (Ours) | 66.68 | 13.45 | 0.17 |

Table 2: Quantitative comparison on abdominal CT dataset. Statistically significant results are presented in bold.

score than the sub-optimal approach MemWarp. The boxplot shows SAMIR outperforms traditional methods in overall and per-structure Dice scores.

Ablation Study

To evaluate the contribution of each module, we conducted an ablation study on the abdominal dataset, and the results are shown in Table 3. The results show that both the SAM

| Modules | SAM encoder | HFC loss | Dice (%) ↑ | HD95(mm) ↓ | SDlogJ ↓ |
|---------|-------------|----------|--------------|------------|----------|
| | | | | | |
| | ✓ | ✓ | 64.90 | 15.86 | 0.17 |
| | | ✓ | 66.19 | 13.93 | 0.17 |
| | | ✓ | 66.68 | 13.45 | 0.17 |

Table 3: Evaluation of the effectiveness of the SAM encoder and HFC loss used in SAMIR on the abdominal CT dataset.

encoder and HFC loss enhance registration performance. This outcome confirms the effectiveness of SAM’s structure-aware visual features and validates the multi-scale feature alignment capability of the HFC loss.

Different Versions of SAM. SAM offered three backbones with varying parameter sizes: ViT-B, ViT-L, and ViT-H. To assess model scale impact, we compared these backbones on the ACDC dataset, as shown in Table 4. Experiments showed that ViT-L and ViT-H outperformed ViT-B in Dice coefficients, but they share a similar registration performance. This may be because ViT-L already captures sufficient structure information for subsequent registration, and further parameter increases provided limited benefit. In addition, we also compared with a fine-tuned ViT-B model released in MedSAM(Ma et al. 2024), which was trained

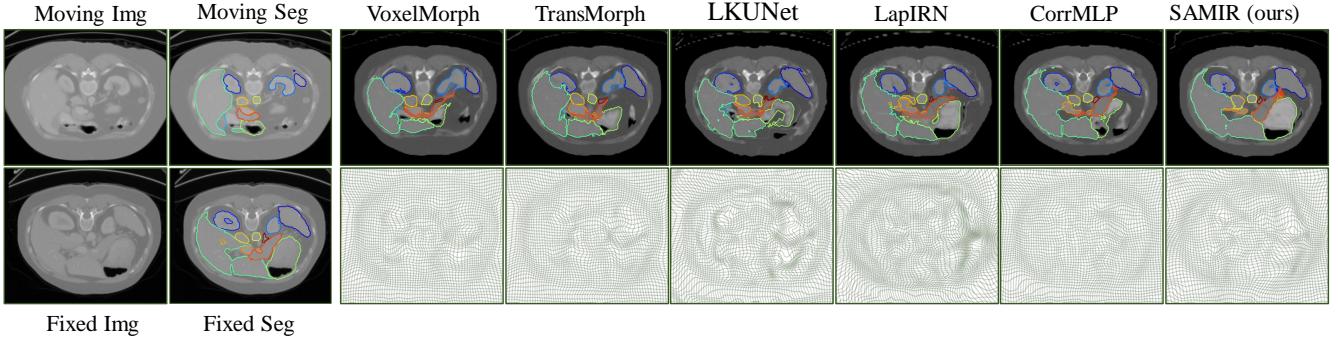


Figure 4: Visual comparison between our SAMIR and SOTA methods on the abdomen CT dataset.

| Model | Dice (%) \uparrow | HD95 (mm) \downarrow | SDlogJ \downarrow |
|--------------|---------------------|------------------------|---------------------|
| SAM-ViT B | 80.49 | 8.39 | 0.046 |
| SAM-ViT L | 80.74 | 8.27 | 0.047 |
| SAM-ViT H | 80.74 | 8.22 | 0.048 |
| MedSAM-ViT B | 80.97 | 7.90 | 0.073 |

Table 4: Comparison of the performance of different SAM models on the ACDC dataset.

| k | Dice (%) \uparrow | HD95 (mm) \downarrow | SDlogJ \downarrow |
|------|---------------------|------------------------|---------------------|
| 256 | 65.36 | 14.45 | 0.17 |
| 512 | 66.68 | 13.45 | 0.17 |
| 1024 | 67.05 | 13.36 | 0.17 |

Table 5: Comparison of dice with different input sizes for the SAM encoder on Abdomen CT dataset.

on 1,570,263 medical image masks covering 10 imaging modalities. We observed some improvement over the original model, but the gain was limited($p > 0.05$). This indicates that the 3D Head has learned sufficient medical domain characteristics of the target and adapted to the downstream task. Moreover, our framework can be easily integrated with other large models, suggesting further potential for performance improvement.

Different Input Sizes for the SAM Encoder. To determine the optimal input image size $k \times k$ for the SAM encoder, we conducted ablation experiments under various resolution settings shown in Table 5. The results show that for abdominal image registration, the model achieves the best performance when the input is upsampled to $k = 1024$. Given the high variability in organ size and morphology within the abdominal cavity, higher spatial resolution helps enhance the visibility of smaller structures (e.g. left and right adrenal gland), leading to more accurate alignment. However, since the performance improvement was not significant ($p > 0.05$) and the computational cost increased substantially, we ultimately selected $k = 512$.

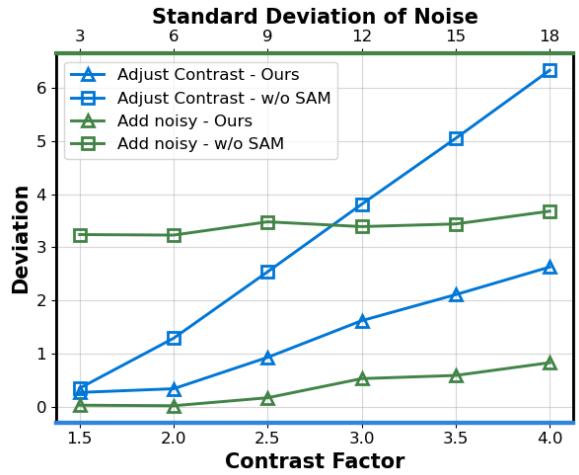


Figure 5: Registration performance of SAMIR under different interference.

Robustness Evaluation

Our SAMIR utilizes structure-aware features rather than raw images for registration, potentially offering improved interference robustness over traditional image-only approaches. To evaluate the model’s robustness, we conducted experiments by adjusting the contrast through gamma correction and adding Gaussian noise with varying standard deviations to the input data, as shown in Figure 5. Analysis reveals that SAMIR achieves superior performance in handling noise and intensity variance across moving and fixed images, which can be attributed to the discriminative structure-aware features learned by the SAM encoder.

Conclusion

This paper proposed a feature-driven registration framework, SAMIR, via efficient feature learning based on the SAM encoder. A SAFE block composing the SAM encoder is designed to extract efficient modality-invariant features from the raw images, with a PDFP block to achieve coarse-to-fine registration, in order to handle large deformation. The HFC loss is used to further enhance the anatomical structure consistency on a multi-scale. Experimental results

demonstrate that SAMIR, through its efficient feature learning mechanism, achieves significantly superior registration accuracy compared to SOTA methods while exhibiting enhanced robustness against intensity variations and noise interference. Future research directions include improving inter-slice correlation in SAM feature extraction and extending the framework to multi-modality registration scenarios.

References

- Balakrishnan, G.; Zhao, A.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2018. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9252–9260.
- Balakrishnan, G.; Zhao, A.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2019. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8): 1788–1800.
- Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; Yang, X.; Heng, P.-A.; Cetin, I.; Lekadir, K.; Camara, O.; Ballester, M. A. G.; et al. 2018. Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Transactions on Medical Imaging*, 37(11): 2514–2525.
- Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.
- Chen, C.; Miao, J.; Wu, D.; Zhong, A.; Yan, Z.; Kim, S.; Hu, J.; Liu, Z.; Sun, L.; Li, X.; et al. 2024a. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98: 103310.
- Chen, J.; Frey, E. C.; He, Y.; Segars, W. P.; Li, Y.; and Du, Y. 2022a. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82: 102615.
- Chen, X.; Diaz-Pinto, A.; Ravikumar, N.; and Frangi, A. F. 2021a. Deep learning in medical image registration. *Progress in Biomedical Engineering*, 3(1): 012003.
- Chen, X.; Hu, R.; Liu, M.; Zhang, Y.; Yaonan, W.; and Zhang, H. 2024b. Encoder-Only Image Registration. *arXiv:arXiv:2509.00451*.
- Chen, X.; Liu, M.; Wang, R.; Hu, R.; Liu, D.; Li, G.; and Zhang, H. 2024c. Spatially covariant image registration with text prompts. *IEEE Transactions on Neural Networks and Learning Systems*, 1–11.
- Chen, X.; Xia, Y.; Ravikumar, N.; and Frangi, A. F. 2021b. A deep discontinuity-preserving image registration network. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 46–55. Springer.
- Chen, X.; Xia, Y.; Ravikumar, N.; and Frangi, A. F. 2022b. Joint segmentation and discontinuity-preserving deformable registration: Application to cardiac cine-MR images. *arXiv:2211.13828*.
- Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Sun, L. J. H.; He, J.; Zhang, S.; Zhu, M.; and Qiao, Y. 2023. SAM-Med2D. *arXiv preprint arXiv:2308.16184*.
- Cheng, X.; Zhang, T.; Lu, W.; Meng, Q.; Frangi, A. F.; and Duan, J. 2025. SACB-Net: Spatial-awareness Convolutions for Medical Image Registration. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 5227–5237.
- Dalca, A. V.; Balakrishnan, G.; Guttag, J.; and Sabuncu, M. 2019a. Unsupervised Learning of Probabilistic Diffeomorphic Registration for Images and Surfaces. *Medical Image Analysis*, 57: 226–236.
- Dalca, A. V.; Balakrishnan, G.; Guttag, J.; and Sabuncu, M. R. 2019b. Unsupervised Learning of Probabilistic Diffeomorphic Registration for Images and Surfaces. *Medical Image Analysis*, 57: 226–236.
- Dong, P.; Wang, L.; Lin, W.; Shen, D.; and Wu, G. 2017. Scalable joint segmentation and registration framework for infant brain images. *Neurocomputing*, 229: 54–62.
- Huang, S.; Xu, T.; Shen, Z.; Saeed, S. U.; Yan, W.; Barratt, D.; and Hu, Y. 2024. One registration is worth two segmentations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 665–675.
- Jia, X.; Bartlett, J.; Chen, W.; Song, S.; Zhang, T.; Cheng, X.; Lu, W.; Qiu, Z.; and Duan, J. 2023. Fourier-net: Fast image registration with band-limited deformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1015–1023.
- Jia, X.; Bartlett, J.; Zhang, T.; Lu, W.; Qiu, Z.; and Duan, J. 2022a. U-net vs transformer: Is u-net outdated in medical image registration? In *International Workshop on Machine Learning in Medical Imaging*, 151–160. Springer.
- Jia, X.; Thorley, A.; Chen, W.; Qiu, H.; Shen, L.; Styles, I. B.; Chang, H. J.; Leonardis, A.; de Marvao, A.; O'Regan, D. P.; Rueckert, D.; and Duan, J. 2022b. Learning a Model-Driven Variational Network for Deformable Image Registration. *IEEE Transactions on Medical Imaging*, 41(1): 199–212.
- Kaiming, H.; Xinlei, C.; Saining, X.; Yanghao, L.; Piotr, D.; and Ross, G. 2021. Masked Autoencoders Are Scalable Vision Learners. *arXiv:2111.06377*.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. In *2023 IEEE/CVF International Conference on Computer Vision*, 3992–4003.
- Lee, M. C.; Oktay, O.; Schuh, A.; Schaap, M.; and Glocker, B. 2019. Image-and-spatial transformer networks for structure-guided image registration. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 337–345. Springer.
- Li, Z.; Tian, L.; Mok, T. C.; Bai, X.; Wang, P.; Ge, J.; Zhou, J.; Lu, L.; Ye, X.; Yan, K.; et al. 2023. SAMConvex: Fast Discrete Optimization for CT Registration Using Self-supervised Anatomical Embedding and Correlation Pyramid. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 559–569. Springer.

- Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment Anything in Medical Images. *Nature Communications*, 15: 654.
- Meng, M.; Feng, D.; Bi, L.; and Kim, J. 2024. Correlation-aware Coarse-to-fine MLPs for Deformable Medical Image Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9645–9654.
- Mok, T. C.; and Chung, A. C. 2020a. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 211–221. Springer.
- Mok, T. C. W.; and Chung, A. C. S. 2020b. Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 211–221. Cham: Springer International Publishing. ISBN 978-3-030-59716-0.
- Oquab, M.; Dariseti, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOV2: Learning Robust Visual Features without Supervision.
- Siebert, H.; Hansen, L.; and Heinrich, M. P. 2021. Fast 3D registration with accurate optimisation and little learning for Learn2Reg 2021. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 174–179. Springer.
- Wang, H.; Ni, D.; and Wang, Y. 2024. Recursive Deformable Pyramid Network for Unsupervised Medical Image Registration. *IEEE Transactions on Medical Imaging*, 1–1.
- Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023. SegGPT: Towards Segmenting Everything In Context. In *2023 IEEE/CVF International Conference on Computer Vision*, 1130–1140.
- Wenhui, L.; Wei, X.; Kang, L.; Xiaofan, Z.; and Shaoting, Z. 2025. MedLSAM: Localize and segment anything model for 3D CT images. *Medical Image Analysis*, 99: 103370.
- Xu, H.; Xue, T.; Liu, D.; Chen, Y.; Zhang, F.; Westin, C.-F.; Kikinis, R.; O'Donnell, L. J.; and Cai, W. 2025. MultiCo3D: Multi-Label Voxel Contrast for One-Shot Incremental Segmentation of 3D Neuroimages. In *The 29th International Conference on Information Processing in Medical Imaging (IPMI) 2025*.
- Xu, Z.; Lee, C. P.; Heinrich, M. P.; Modat, M.; Rueckert, D.; Ourselin, S.; Abramson, R. G.; and Landman, B. A. 2016. Evaluation of six registration methods for the human abdomen on clinically acquired CT. *IEEE Transactions on Biomedical Engineering*, 63(8): 1563–1572.
- Young, S. I.; Balbastre, Y.; Dalca, A. V.; Wells, W. M.; Iglesias, J. E.; and Fischl, B. 2022. SuperWarp: Supervised Learning and Warping on U-Net for Invariant Subvoxel-Precise Registration. In *Biomedical Image Registration*, 103–115. Springer International Publishing.
- Yue, W.; Zhang, J.; Hu, K.; Xia, Y.; Luo, J.; and Wang, Z. 2024. SurgicalSAM: efficient class promptable surgical instrument segmentation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press.
- Zhang, H.; Chen, X.; Hu, R.; Liu, D.; Li, G.; and Wang, R. 2024. MemWarp: Discontinuity-Preserving Cardiac Registration with Memorized Anatomical Filters. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 671–681. Springer.
- Zhao, S.; Dong, Y.; Chang, E. I.; Xu, Y.; et al. 2019. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10600–10610.
- Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2022. iBOT: Image BERT Pre-Training with Online Tokenizer. *International Conference on Learning Representations (ICLR)*.