

SAM: Self-supervised Learning of Pixel-wise Anatomical Embeddings in Radiological Images

Ke Yan*, Jinzheng Cai, *Member, IEEE*, Dakai Jin, *Member, IEEE*, Shun Miao, Dazhou Guo, Adam P. Harrison, Youbao Tang, Jing Xiao, Jingjing Lu*, Le Lu, *Fellow, IEEE*

Abstract—Radiological images such as computed tomography (CT) and X-rays render anatomy with intrinsic structures. Being able to reliably locate the same anatomical structure across varying images is a fundamental task in medical image analysis. In principle it is possible to use landmark detection or semantic segmentation for this task, but to work well these require large numbers of labeled data for each anatomical structure and sub-structure of interest. A more universal approach would learn the intrinsic structure from *unlabeled* images. We introduce such an approach, called Self-supervised Anatomical eMbedding (SAM). SAM generates semantic embeddings for each image pixel that describes its anatomical location or body part. To produce such embeddings, we propose a pixel-level contrastive learning framework. A coarse-to-fine strategy ensures both global and local anatomical information are encoded. Negative sample selection strategies are designed to enhance the embedding’s discriminability. Using SAM, one can label any point of interest on a template image and then locate the same body part in other images by simple nearest neighbor searching. We demonstrate the effectiveness of SAM in multiple tasks with 2D and 3D image modalities. On a chest CT dataset with 19 landmarks, SAM outperforms widely-used registration algorithms while only taking 0.23 seconds for inference. On two X-ray datasets, SAM, with only one labeled template image, surpasses supervised methods trained on 50 labeled images. We also apply SAM on whole-body follow-up lesion matching in CT and obtain an accuracy of 91%. SAM can also be applied for improving image registration and initializing CNN weights.

Index Terms—Contrastive learning, follow-up lesion matching, landmark detection, pixel-wise embedding, self-supervised learning.

I. INTRODUCTION

In medical image analysis, many tasks essentially aims at locating certain anatomical structures. For example, landmark detection finds anatomical points that are useful for diagnosis, quantification, therapy evaluation and planning, or as a preprocessing step for other downstream tasks. [1],

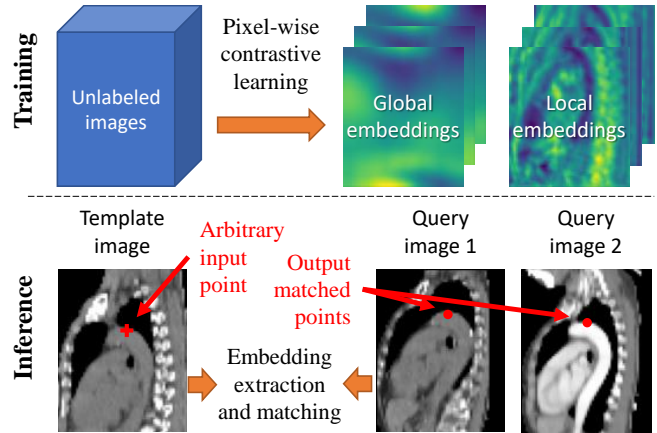


Fig. 1: Self-supervised anatomical embedding (SAM) and its application of anatomical location matching.

[2]. Lesion matching locates follow-up lesions in multiple images, which is an important clinical task for radiologists to longitudinally track disease progress [3]. However, supervised algorithms [2], [3] need sufficient labeled data for training. Annotating medical images is laborious, expensive, and requires considerable expertise [4]. Moreover, supervised models are restricted to only discovering landmarks or lesions that were *a priori* labeled and trained, but many labeled datasets are limited in their coverage and fine-grainedness.

Human organs are intrinsically structured, so there is an inherent consistency underlying their appearance and layout in radiological images such as computed tomography (CT) and X-rays. Intuitively, we can design an algorithm to discover these intrinsic patterns by learning from *unlabeled* images. Given a certain anatomical location (e.g., top of aortic arch) in one image, the trained algorithm should be able to find the corresponding location in other images (see Fig. 1), although it was not trained specifically to detect such landmarks. We aim to develop a universal algorithm that learns from unlabeled radiological images to detect arbitrary points of interest. It can generate embeddings on each image pixel to encode its anatomical context information, so that the same body part in different images express similar embeddings and can be retrieved by simple nearest neighbor searching. In this paper, we use the terms anatomical location or body part to describe the semantic location of a pixel, such as “center of the right

K. Yan is with Alibaba Group DAMO Academy, Beijing, 100026, China; J. Cai, D. Jin and L. Lu are with Alibaba Group DAMO Academy, New York, NY 10014, USA.

A. P. Harrison, Y. Tang, and D. Guo are with PAII Inc., Bethesda, MD 20817, USA. S. Miao is with Xpeng Motors, San Diego, CA, USA.

J. Xiao is with Ping An Insurance (Group) Company of China, Ltd., Shenzhen, 510852, China.

J. Lu is with Beijing United Family Hospital, Beijing, 100015, China. Corresponding authors: Ke Yan (email: yankethu@gmail.com); Jingjing Lu (email: cjr.lujingjing@vip.163.com)

lobe of thyroid” or “femur head”.

Learning anatomical embeddings that are universal across body parts, without anatomical labels in training, is a challenging task. For example, in CT images the algorithm has to memorize the 3D contextual appearance of numerous body parts so as to generate globally distinguishable embeddings. Meanwhile, it needs to encode local information to differentiate adjacent structures with similar appearance for accurate localization. In addition, the embeddings should be robust to the size, shape, intensity, and texture diversity of body parts caused by inter-subject variation, organ deformation, contrast injection, and pathological changes.

In this work, we propose Self-supervised Anatomical eMbedding (SAM) to achieve these goals. We adopt the recent self-supervised contrastive learning [5]–[7] approach and propose a pixel-level contrastive learning framework to differentiate pixels of varying body parts. During training, we use data augmentation to create synthetic body part pairs and simulate appearance changes across different images. To cover both global and local information, we design a coarse-to-fine architecture with two-level embeddings. The global embedding is trained to distinguish every body part on a coarse scale, helping the local embedding to focus on a smaller region to differentiate with finer features. The selection of negative samples (in our case, dissimilar body parts) is a crucial step of contrastive learning. We design several strategies to carefully control the difficulty and diversity of the selected negative samples. Finally, during inference, embeddings can be matched efficiently using 2D or 3D convolutional operations.

Due to its unsupervised and universal nature, SAM can be trained easily and applied to various tasks. A natural application is “one-shot” landmark detection with one labeled template image. We conduct such experiments on a chest CT dataset and two X-ray datasets (pelvis and hand) [8]. On the chest CT dataset, SAM outperforms widely-used deformable registration algorithms [9]–[12] while only taking 0.23 seconds to infer an image. On the X-ray datasets, our method with one labeled template outperforms the supervised HRNet [13] trained on 50 labeled images. Another application is lesion matching [14], [15]. On 3,556 lesion pairs manually annotated in the DeepLesion dataset [14], [16], SAM outperforms previous supervised matching algorithms [3], [14], [17] (91% versus 82% in accuracy). SAM can also be applied to improve unsupervised deformable image registration by providing corresponding points between images, semantic features, and losses. Our SAM-enhanced registration algorithm [18] achieves superior accuracy compared to existing learning-based [12] or optimization-based methods [9]–[11].

Our major contributions are three-fold: **1)** The problem of universal anatomical embedding learning is tackled for the first time. We demonstrate that discriminative and robust embeddings can be learned in a self-supervised manner. **2)** A pixel-level contrastive learning framework is proposed with a coarse-to-fine architecture and customized negative sampling strategies. The proposed algorithm is easy to train and fast. **3)** Most importantly, we show the learned embeddings are widely applicable in various applications (landmark detection, lesion matching, registration, etc.) on different image modalities (3D

CT and 2D X-ray) of varying body parts (chest, hand, pelvis, etc.). Similar or superior performance is achieved compared to existing registration or supervised algorithms.

II. RELATED WORK

Self-supervised learning (SSL). In SSL, visual features are learned from unlabeled data by optimizing a heuristic pretext task, such as predicting image rotations [19] and inpainting [20]. Contrastive learning algorithms [5]–[7] use a simple paradigm of data augmentation and image instance discrimination and achieve impressive performance. Recently, some SSL methods were developed to learn dense representations, e.g., VADeR [21] and DenseCL [22]. They focus on network pretraining and need finetuning on downstream tasks such as object detection and segmentation. However, our goal is learning anatomical embeddings, which is coherent in both training and inference; thus, no finetuning is needed. Both VADeR and DenseCL are designed for natural images and rely on image-wise contrast as initialization [21] or as an additional loss [22], which does not apply to radiological images. This is because each radiological image contains the same set of body parts, so it is hard to define image-wise negative samples as all images are similar. Our method is specially designed for radiological images with strategies crucial to make the embeddings effectively discriminative and fine-grained, i.e., a coarse-to-fine structure and negative sampling algorithms.

Radiological images are suitable for SSL because of the intrinsic structure of the human anatomy. Related studies can be classified into three major categories. Content restoration algorithms [24]–[26] reconstruct original images from corrupted or cropped ones, aiming to pretrain network weights. For instance, Models Genesis [24] designs a series of image corruption strategies. It outperforms training from scratch in a wide range of downstream applications covering both segmentation and classification. Transformation prediction algorithms first do data augmentation to the image, and then predict the augmentation parameters, such as translation [27], patch order [28], and rotation [29]. An example is Rubik’s cube [28], which uses cube rearrangement and cube rotation as proxy tasks to enforce networks to learn translational and rotational invariant features from raw 3D data. Image-wise [30], patch-wise [31] and superpixel-wise [32] contrastive learning have also been proposed to pretrain networks [30] or learn representations for few-shot image segmentation [31], [32]. None of these works are designed for universal and fine-grained anatomical point matching. In this task, each pixel must be encoded to be distinguishable from other body parts, which is SAM’s explicit training objective. Other self-supervised algorithms use different proxy tasks that are not directly applicable in this task. Additionally, methods in [31] and [32] require relatively complex multi-stage training or superpixel extraction, while our method is simple and end-to-end trainable.

Anatomy-related learning. Anatomical landmark detection can be used to initialize or aid many medical imaging tasks [8], [33]. Heatmap and coordinate-based methods have been studied [8]. Deep adaptive graph was proposed in [8] to

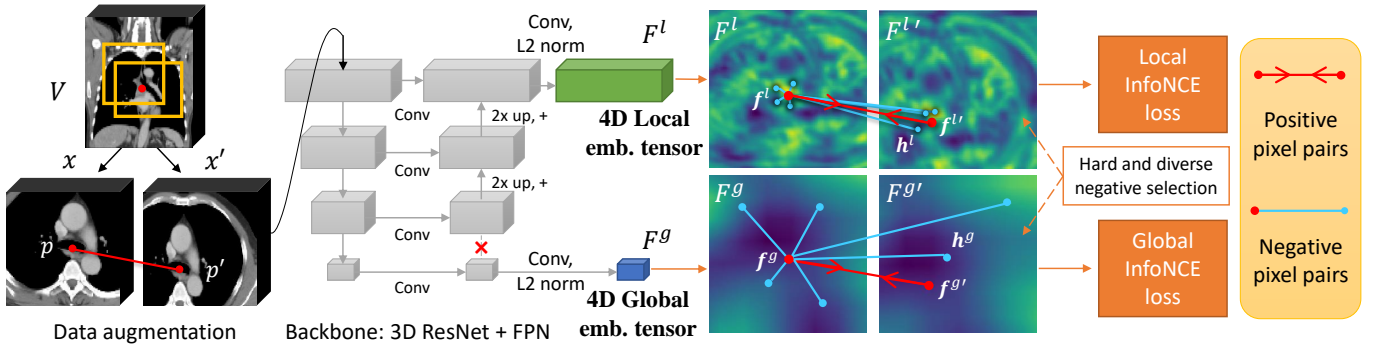


Fig. 2: The learning process of SAM. We first do random data augmentation to an unlabeled image, then send the two transformed patches to a network to generate a global and a local embedding vector for each pixel (Sec. III-A). One embedding channel of an axial image slice is visualized in the figure with the global embedding zoomed in. Next, we use a global and a local InfoNCE loss [23] to encourage matched pixels (positive pairs) to have similar embeddings, while push unmatched ones (negative pixels) apart (Sec. III-B). The negative pixels are random pixels in the two patches guided by hard and diverse sample selection strategies (Sec. III-C).

capture relationships among landmarks and achieved state-of-the-art results in three X-ray datasets. However, these methods all need training labels and can only predict the predefined anatomical points, whereas our method is unsupervised and uses embedding matching to find arbitrary landmarks of interests. Anatomy-specific classification [34] or regression [35] methods can predict predefined body part classes or scores of CT slices. Compared to their slice-wise coarse predictions, our pixel-wise embedding is more fine-grained.

Image registration is popularly used to find pixel-wise correspondences in medical images. Traditional algorithms [9]–[11] iteratively optimize a predefined objective function. Recent deep learning-based methods [12], [36] directly predict a deformation field to align an image pair using the spatial transformer network [37]. They are fast in inference but often less accurate than state-of-the-art traditional algorithms [10], [38]. Registration needs an image pair as input and aims to align each pixel via fundamental appearance-based similarity measures. By contrast, SAM tackles the pixel matching problem in a different perspective by semantically representing each pixel given one single image. Its core application is anatomical point matching instead of full-volume registration, but we will show it can be readily used to improve existing registration algorithms [18].

III. METHOD

The objective of self-supervised anatomical embedding (SAM) is to encode the semantic anatomical information of each pixel, so that similar body parts in different images have similar embeddings. Inspired by previous works on contrastive learning [5]–[7], we propose a pixel-level contrastive learning framework as shown in Fig. 2. Essentially, we let the model contrast the appearance of random pixels, so that it can gradually discover distinctive patterns of each anatomical point by itself. The proposed framework comprises the following steps: stochastic data augmentation for training patch generation, coarse-to-fine CNN for pixel-wise feature embedding, and positive & negative sample selection for contrastive loss

computation. In this section, we elaborate our framework for 3D images such as CT. It is straightforward to adapt SAM to 2D images such as X-ray by changing the network from 3D to 2D.

A. Coarse-to-Fine Network Architecture

As shown in Fig. 2, in training, we first take an unlabeled CT volume V , then crop two 3D patches with random location and size, and resize them to the same shape, namely $x, x' \in \mathbb{R}^{d \times h \times w}$. The patches are sent to a fully-convolutional network to extract pixel-wise embeddings. To learn universal anatomical embeddings, on the one hand, SAM needs to memorize the 3D contextual appearance of each body part so its embedding is distinguishable from all other body parts. On the other hand, it needs to encode local information to differentiate adjacent structures with similar appearance for accurate localization. To get the best of both worlds, we propose a coarse-to-fine architecture that predicts a global embedding and a local one for each pixel. A lightweight 3D ResNet-18 [39] is adopted as backbone with a feature pyramid network (FPN) [40] to fuse multi-scale features. The ResNet-18 is initialized with ImageNet pretrained weights using the inflated 3D (I3D) technique¹ [41]. The 4D global embedding tensor F^g and the local one F^l are both generated from the FPN features using separate $3 \times 3 \times 3$ convolutions and L2 normalization layers. F^g is from the coarsest FPN level with a larger stride and more abstract features, while F^l is from the finest FPN level with a smaller stride and detailed features. Specifically, we cut the connection between the coarsest FPN feature and its upper level to make F^g and F^l more independent, see Fig. 2. Examples of the learned F^g and F^l can be found in Figs. 2 and 1, where F^l contains more high-frequency details.

¹Suppose a 3D filter is $s \times s \times d$, I3D essentially repeats a pretrained $s \times s$ 2D filter in the z -dimension d times, and then rescales the obtained 3D filter by dividing by d .

B. Pixel Pairs and Loss Function

In this step, we pick matched and unmatched pixel pairs to train SAM. Note that although we crop two patches x, x' from the original volume V , we aim to contrast the *pixel* inside the patches, which is different from previous works that contrast patches [6], [31], [42]. When the two patches overlap, we can find a pixel pair $p \in x, p' \in x'$ from them so that p and p' actually correspond to the same pixel in V , see Fig. 2. They are treated as a positive pair. We randomly sample n_{pos} pixel pairs from the overlapping area. The sampled pixels do not need to be close to patch centers, because SAM needs to learn to embed all pixels in an image, including those at the center and borders. The positive embeddings at p_i, p'_i are $\mathbf{f}_i, \mathbf{f}'_i \in \mathbb{R}^c, 1 \leq i \leq n_{\text{pos}}$. For each positive pair (p_i, p'_i) , we also find n_{neg} pixels as negative pixels. They are randomly sampled from x and x' as long as their distance from p_i and p'_i is larger than δ ($\delta = 3\text{mm}$ in this paper). We denote the embeddings at the negative pixels of pair i as $\mathbf{h}_{ij} \in \mathbb{R}^c, 1 \leq j \leq n_{\text{neg}}, j$ being the index of negative pixels. The InfoNCE loss [23] for pixels in this patch pair is defined as

$$\mathcal{L} = - \sum_{i=1}^{n_{\text{pos}}} \log \frac{\exp(\mathbf{f}_i \cdot \mathbf{f}'_i / \tau)}{\exp(\mathbf{f}_i \cdot \mathbf{f}'_i / \tau) + \sum_{j=1}^{n_{\text{neg}}} \exp(\mathbf{f}_i \cdot \mathbf{h}_{ij} / \tau)}, \quad (1)$$

where $\tau = 0.5$ is a temperature parameter and \cdot is the inner product operation. We call \mathbf{f}_i anchor embeddings. Since we want (1) to be symmetric with \mathbf{f}_i and \mathbf{f}'_i [6], another loss term with swapped \mathbf{f}_i and \mathbf{f}'_i is also computed and added. Eq. (1) is applied to both global and local embeddings.

In summary, we use data augmentation to simulate appearance changes across different image volumes. Given an anchor pixel $p \in x$, its matched pixel $p' \in x'$ represents the same anatomical location in another (simulated) image; Its unmatched pixels are in different anatomical locations that are not physically adjacent to it. Eq. (1) pulls the matched pixels (positive pairs) together and pushes unmatched ones (negative pairs) apart in the embedding space. This strategy is simple yet effective. It can make the network learn to distinguish different body parts effectively with the help of the negative sampling strategy described in the next section.

C. Hard and Diverse Negative Sampling

We propose several strategies to select hard and diverse negative pixels and make global and local embeddings focus on different scopes. The global embedding F^g is responsible for differentiating all body parts on a coarse scale. To make it more discriminative, we perform online hard negative selection [43] to train it. For each global anchor embedding \mathbf{f}_i^g , we compute its cosine similarity map with the global embedding tensors F^g and $F^{g'}$ to get similarity maps S_i^g and $S_i^{g'}$. Then, hard negatives \mathbf{h}_{ij}^g are selected to be the n_{neg} embeddings with the greatest similarity with \mathbf{f}_i^g (excluding \mathbf{f}_i^g and $\mathbf{f}_i^{g'}$ themselves). Additionally, we further populate the negative samples to contain more diverse body parts. For each \mathbf{f}_i^g , we randomly sample an additional n_{rand}^g pixels from all patches across all images within a training batch, except the two patches of the image that \mathbf{f}_i^g is on. Because these pixels

originate from different images it is very unlikely the sampled pixels are in the same semantic position as \mathbf{f}_i^g .

Moving on to the local embedding, it uses the same positive pixel pairs (p_i, p'_i) as the global version, but with different negative samples. For each local anchor \mathbf{f}_i^l , we first compute its local similarity maps $S_i^l, S_i^{l'}$ with $F^l, F^{l'}$, then upsample $S_i^g, S_i^{g'}$ to be the same size as S_i^l , and finally use the combined global and local similarity maps $S_i^g + S_i^l, S_i^{g'} + S_i^{l'}$ to select hard negatives pixels for \mathbf{f}_i^l . Examples of S^l and upsampled S^g can be found in Figs. 3 and 7. The definition of hard negative pixel is similar to that in global embeddings. The pixels with higher values on the combined similarity map are harder except the positive pixel pair themselves. When the global embedding converges, S_i^g will be high only in the neighborhood of p_i , so that hard negatives can be selected mostly in the local area. During inference we use the peak of $S^g + S^l$ to find matched points, so using $S^g + S^l$ for local hard negative selection is consistent with our inference process. Meanwhile, because the local embedding tensor has a high resolution, the selected local hard negatives will mostly be adjacent pixels and correlated. To improve diversity, we first find the top $n_{\text{cand}}^l > n_{\text{neg}}$ hard negative candidates, then randomly sample n_{neg} embeddings from them to be used in Eq. (1). Algorithm 1 summarizes the complete training procedure of SAM.

Algorithm 1 Self-supervised anatomical embedding (SAM)

Input: Unlabeled images $\{V\}$; batch size b , pixel sampling parameters $n_{\text{pos}}, n_{\text{neg}}, n_{\text{rand}}^g, n_{\text{cand}}^l$.

Output: Trained network.

- 1: In each training batch, randomly select b images from $\{V\}$.
 - 2: **for** each image **do**
 - 3: Run random data augmentation to get a patch pair (x, x') .
 - 4: Compute global and local embedding tensors of the two patches, $F^g, F^{g'}, F^l, F^{l'}$.
 - 5: Sample positive pixel pairs (p_i, p'_i) from the overlapping area of x and x' , then extract global and local positive embeddings pairs: $(\mathbf{f}_i^g, \mathbf{f}_i^{g'}), (\mathbf{f}_i^l, \mathbf{f}_i^{l'}), 1 \leq i \leq n_{\text{pos}}$.
 - 6: **for** each positive pair i **do**
 - 7: Compute similarity maps $S_i^g, S_i^{g'}, S_i^l, S_i^{l'}$.
 - 8: Find global hard negatives $\mathbf{h}_{ij}^g, 1 \leq j \leq n_{\text{neg}}$.
 - 9: Sample global random negatives $\mathbf{h}_{ik}^g, 1 \leq k \leq n_{\text{rand}}^g$ from the other $b - 1$ patch pairs in the training batch.
 - 10: Pool the hard and random negatives to get the final global negatives. The variable n_{neg} in Eq. 1 now becomes $n_{\text{neg}} + n_{\text{rand}}^g$ for global embeddings.
 - 11: Find local hard negatives $\mathbf{h}_{ij}^l, 1 \leq j \leq n_{\text{cand}}^l$. Randomly sample n_{neg} from them as the final local negatives. The variable n_{neg} in Eq. 1 is unchanged for local embeddings.
 - 12: **end for**
 - 13: Compute the global and local InfoNCE loss $\mathcal{L}^g, \mathcal{L}^l$ in Eq. 1. The final loss is $\mathcal{L} = \mathcal{L}^g + \mathcal{L}^l$.
 - 14: **end for**
-

D. Application: Anatomical Point Matching

Fig. 3 depicts the inference procedure of SAM for anatomical point matching. To locate a certain point of interest, we first need to label it on a template image (also known as an atlas [44] or a support image [32]). Then, given an unlabeled query image, we compute the global and local embedding tensors. After extracting the anchor embedding vectors from

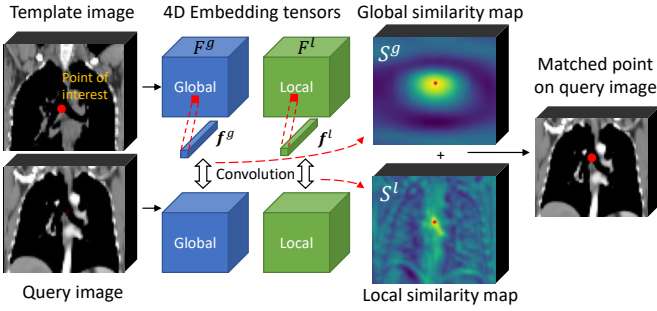


Fig. 3: The inference procedure of SAM for anatomical point matching. Red dots are the point of interest or matched point.

the point of interest of the template, we compute the similarity maps S^g, S^l between the anchor vectors and the query tensors. The embeddings are L2-normalized, so the cosine similarity maps can be efficiently computed using the convolutional operation on GPU. Finally, we upsample S^g, S^l to the size of the original image, and find the peak of $S^g + S^l$ as the matched anatomical point.

E. Application: Enhancing Image Registration

SAM is designed to match sparse points. It may be possible to directly use it for full-volume registration by matching every pixel between the fixed image and the moving image, but it is highly inefficient as there are millions of pixels in a 3D CT scan. We have proposed a SAM-Enhanced deformable registration algorithm called SAME [18]. It consists of three steps. (1) SAM-affine. We first sample a sparse grid on the fixed image and discard points outside the body, then use SAM to match their corresponding points on the moving image. The matches with similarity scores higher than a threshold are kept. With these point pairs, an affine transformation matrix can be estimated by simple least squares fitting. (2) SAM-coarse, which uses SAM to compute a new correspondence grid on the transformed images to interpolate a coarse-level deformation field. These two steps are efficient, require no additional training, and can provide a good initialization for the final step. (3) Lastly, SAM-VoxelMorph, which enhances the deep learning-based VoxelMorph registration method [12] by incorporating SAM-based correlation features [45] and an additional SAM-based similarity loss. More details can be found in our published work [18]. SAME is unsupervised and only needs a trained SAM model.

IV. EXPERIMENTS

We evaluate SAM on four diverse tasks: 3D landmark detection on chest CT, 2D landmark detection on hand and pelvic X-rays, and 3D universal lesion matching on CT. Our self-supervised method is compared with widely-used registration methods [9]–[12], supervised landmark detection [8], [13], [46] and lesion matching [3], [14], [17] methods, **and other feature extraction [47] and network pretraining methods [24], [48].** We will briefly describe our results on 3D CT registration and direct readers to [18] for more details. **As extensions of SAM, we further explore using SAM for network initialization and training SAM with auxiliary label information.**

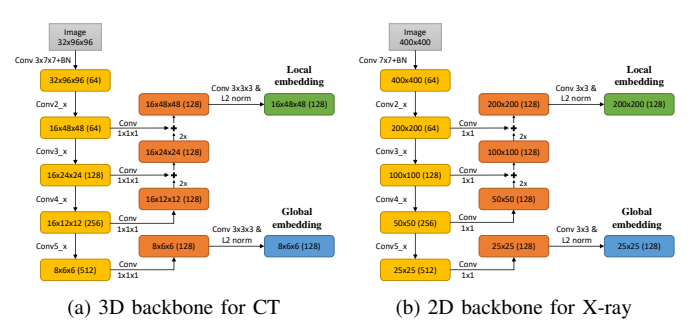


Fig. 4: Detailed coarse-to-fine network structures. The shapes of images and feature maps in training are shown in the format of $z \times y \times x$ (*channel*). The backbones are based on (a) 3D ResNet-18 and (b) 2D ResNet-34 [39]. **Downsampling is done by max pooling in Conv2_x and by convolution with stride in other blocks.** To enlarge the resolution of the feature maps, we use stride 1 in the first convolutional layer and set the z stride to 1 in Conv3_x and Conv4_x. Please zoom in.

A. Tasks and Datasets

For the CT-based tasks, we train a unified SAM model using the combination of three CT datasets: DeepLesion, NIH-Lymph Node (NIH-LN), and an in-house chest CT dataset (ChestCT). DeepLesion [14], [16] is a large CT dataset with 20K sub-volumes covering the whole body. It contains a variety of contrast phases, reconstruction protocols, and pathology. NIH-LN [49] includes 176 chest-abdomen-pelvis CT scans. ChestCT contains 94 patients, each with a contrast-enhanced (CE) and a non-contrast (NC) scan that are pre-aligned. When training SAM, no labels from the three datasets are used. We follow the official data split for DeepLesion, while use all images in NIH-LN for training. For ChestCT, we randomly split the patients to 65 training, 10 validation, and 19 testing.

3D landmark detection is tested using ChestCT, which contains manual annotations of 35 organs for each patient. We calculated 19 landmarks to evaluate SAM, such as the top of left lung, the bifurcation of trachea, etc. To detect landmarks in the one-shot setting, we need to select a template image from the training set. The sample whose landmarks are closest to the average normalized landmarks is chosen as the template [50].

2D landmark detection is conducted on X-rays following the settings in [8]. The hand and pelvic models are trained separately. The hand X-ray dataset contains 378 images for training and 93 for testing. The images resolutions are in $\sim 1500 \times 2000$. 30 landmarks are manually labeled in each image. The pelvic X-rays are taken on each patient’s pelvic bone with resolutions of $\sim 2500 \times 2000$. 800 images are used for training and 200 for testing. The template images are chosen in the same way as in ChestCT. In both hand and pelvic X-rays, there are high structural and appearance variations, caused by various poses and pathologies. Landmarks in X-rays are useful in bone age estimation, arthritis diagnosis, etc.

3D Lesion matching is tested using DeepLesion, which includes 32K lesions annotated on 4K patients. Most patients have multiple follow-up CTs. Given a lesion annotation on one CT, the task of lesion matching is to locate the corresponding lesion on another CT of the same patient. We manually

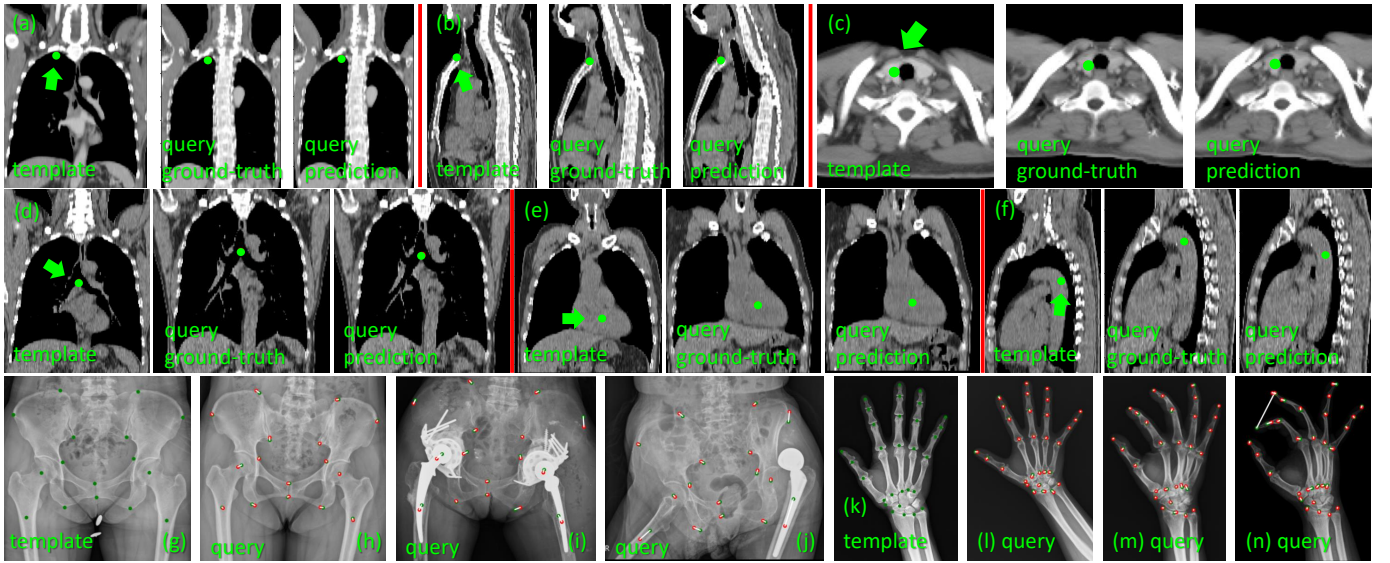


Fig. 5: Examples of landmark detection results of SAM. Each image type (chest CT, pelvic X-ray, and hand X-ray) has only one template image. For CT images (a–f), the views (axial, coronal, sagittal) are selected in order to clearly show the point. For X-ray images (g–n), the green and red points are respectively ground-truth and predicted points, while the white lines show their correspondence. (f), (j), and (n) are failure cases in which the prediction errors are larger.

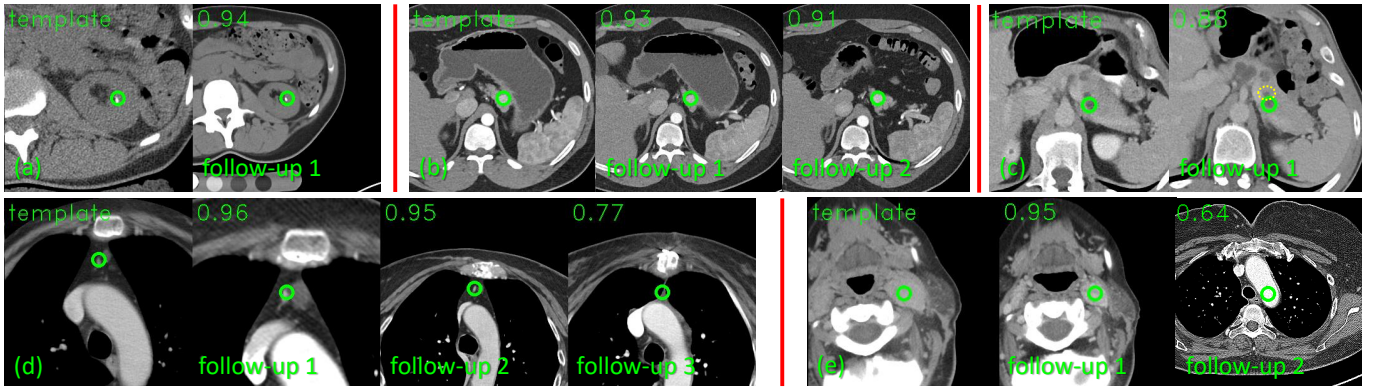


Fig. 6: Examples of lesion matching results of SAM on DeepLesion. Green circles mark the lesion’s central point in the template scan and the detected points in follow-up scans. The predicted similarity scores are also shown. See text for details.

annotate 3,556 lesion pairs on the validation and test sets of DeepLesion. We extract SAM embedding from the center of the box of the template lesion, and use it to detect a matched point in the follow-up scan.

Evaluation metrics: We use the mean radial error (MRE) [8] and maximum radial error as the evaluation metrics. Radial errors are the Euclidean distances of predicted and ground-truth coordinates. For lesion matching, the central points of the lesions’ 3D bounding-boxes are used as landmarks to compute errors. Besides, if the matched point falls in the lesion’s bounding-box, it is regarded as a correct match to calculate an accuracy.

B. Implementation Details

SAM is implemented in PyTorch with the help of data augmentation libraries [51], [52]. It is trained using rectified Adam (RAdam) [53] with a learning rate of 0.0001. The number of patch pairs per batch is 8 for CT and 16 for X-ray.

All models converge in about 25K iterations. We normalize CT images to an isotropic spacing of 2mm, whereas the X-ray images are resized to 512×512 as in [8]. Following [8], the detected landmarks on resized X-rays are mapped back to match the original image size, and then errors are computed. During training, the cropped patch size is $32 \times 96 \times 96$ for CT and 400×400 for X-ray. During inference, the entire image without cropping is sent to the network. We have verified that the deviation in input size during training and inference does not affect performance. During training, suppose the side length of the training patch is b , the side length of the overlapped region between two patches is a . We make the overlap ratio a/b to be a random value between 0.2 and 1. To promote reproducibility, we illustrate the detailed 3D and 2D network architectures in Fig. 4. The backbone network is 3D ResNet-18 for CT and 2D ResNet-34 for X-ray. Further increasing the network depth did not result in significant improvements. The global and local embedding

dimensions are both $c = 128$. The default sampling hyper-parameters for CT are $n_{\text{pos}} = 100$, $n_{\text{neg}} = 500$, $n_{\text{rand}}^g = 1000$, and $n_{\text{cand}}^l = 2 \times 10^4$. For X-ray, they are the same except for $n_{\text{cand}}^l = 5000$ as we hope the local embedding to be more locally focused in 2D images. We will study the selection of the hyper-parameters and show that SAM is robust to their values in Sec. IV-E.

C. Qualitative Results

As shown in Fig. 5, SAM is able to accurately detect various landmarks in different body parts with only one template image. A failure case is shown in (f), where the detection of the top of the descending aorta is off in the z -axis, which is mainly because the tube-shaped aorta does not have discriminative textures along the z -axis. For X-ray images, SAM can locate the landmarks in the presence of body rotation, deformation, and metal prostheses. Failure cases occur when there are severe pathological deformations that significantly differ from the template image (subplots (j) and (n)). More examples are exhibited in the supplementary video. In the video, we click on random anatomical points on one CT image and apply SAM in real-time to find the corresponding points on another CT. The matching results are promising despite the fact that the images are from datasets that SAM has never seen.

SAM can match a variety of lesions effectively in follow-up CTs, see Fig. 6. In (a), it locates a tiny kidney lesion despite the different noise levels in two images. In (b), SAM detects the lesion in “follow-up 2” even though the shape of its surrounding organ has changed. In (c), the algorithm successfully differentiates the true matched lesion from an adjacent one in the dashed yellow circle. (d) is an interesting case where the lesion disappeared in “follow-up 3”, possibly due to surgery. SAM can locate the correct position even if the lesion no longer exists, showing that it did not simply match the lesion texture, but has learned the body part from the anatomical context. Besides, “follow-up 1” in (d) is a CT scan with a limited field of view, which SAM handled well. Sometimes a lesion in the template scan may not exist in the follow-up scan due to the scanning range differences in the z -axis, which is the case in “follow-up 2” of (e). SAM detected an unrelated point with a low similarity score, which can be filtered with a score threshold when used in practice.

D. Quantitative Results

In this section, we compare the performance of SAM with existing registration, landmark detection, and lesion matching algorithms. We also display the results of fully supervised methods trained on all labeled samples as an upper bound for each task. Note that the fully supervised methods should not be directly compared with SAM since they use significantly more training labels on predefined landmarks.

For the 3D landmark detection task, we compare with four widely used conventional registration methods [9]–[11], [54] and the unsupervised deep learning-based method VoxelMorph [12] in one-shot (single atlas) setting. VoxelMorph is initialized using affine registration in Elastix [54]. The template CT image in the training set is registered respectively

TABLE I: Comparison of methods on the ChestCT dataset.

	MRE _{CE}	Max _{CE}	MRE _{NC}	Max _{NC}	Time (s)
Affine [54]	8.4±5.2	32.9	8.5±5.3	33.1	6.82
FFD [11]	5.4±4.1	29.1	5.5±3.9	24.6	86.66
SyN [9]	6.0±5.2	35.1	6.0±4.7	29.3	78.58
DEEDS [10]	4.6±3.3	18.8	4.7±3.4	24.4	50.16
VoxelMorph [12]	7.3±3.6	20.1	7.4±3.7	20.2	7.43
SAM	4.3±3.0	16.4	4.5±3.0	18.5	0.23
SCN (supervised) [46]	2.6±2.2	15.9	2.7±2.3	16.0	4.89

CE: contrast-enhanced; NC: non-contrast; Template and query images are in the same modality (CE or NC).

MRE: mean radial error \pm std.; Max: maximum radial error. Errors are calculated in pixels. The mean, std., and maximum values are computed over results of all landmarks on all test images.

to the images in the test set. The landmark detection errors are shown in Table I. Free-Form Deformation (FFD) [11] and Symmetric Normalization (SyN) [9] are the top-performing methods in registering chest CTs [55], while DEEDS [10] performs the best in a recent survey [56] and is shown to be more accurate than a recent deep learning-based approach [38]. SAM consistently outperforms all these methods in all metrics while taking only 0.23s to process one volume (including embedding extraction and matching on GPU). VoxelMorph performs better than Affine but inferior to some other deformable conventional methods, which is consistent to the findings in [38], [57], [58]. Deformable registration methods require a robust initial alignment to perform well, and may struggle when the two images exhibit large differences in terms of body size/pose, organ shape and respiratory status [56]. In contrast, SAM is able to encode semantic and fine-grained anatomical information, which addresses the landmark matching problem effectively from a novel perspective. We further use a CE template to detect landmarks in NC images, and vice versa. SAM’s MREs are 4.5 and 4.8 (DEEDS: 4.7, 4.7). They are comparable to the results using the same modality to train and test (4.5 and 4.3), showing SAM’s robustness to contrast changes in CT. We have also run paired t -test between SAM and other methods on the CE task. SAM is significantly better than Affine, FFD, SyN, and VoxelMorph ($p < 10^{-9}$) while comparable with DEEDS ($p = 0.05$) despite being much faster. In addition, SAM also has smaller standard deviation and maximum radial error than other methods, indicating that it is more stable and works better in worst cases.

For the 2D X-ray landmark detection tasks, we compare with a strong supervised baseline HRNet [13] and two state-of-the-art methods, Deep Adaptive Graph (DAG) [8] and SpatialConfiguration-Net (SCN) [46]. We test both few-shot and fully supervised settings. In Table II, SAM outperforms HRNet trained on 50 samples (HRNet-50), and DAG-5 in average error. One possible concern is the relatively large variance, which is mainly caused by large errors of a few mismatched outliers (Fig. 5 (j)(n)) and the large size of X-ray images (~ 2000 px). To verify the significance of the results, we conduct paired t -test between SAM and other methods. In hand X-rays, SAM is significantly better than

TABLE II: Results on the X-ray datasets. Supervised methods are trained in few-shot and fully supervised settings (378 and 800 for hand and pelvic datasets, respectively).

Method	# Labeled samples	Hand MRE	Pelvic MRE
HRNet [13]	5	43.3±112.0	139.1±327.5
	50	14.1±14.7	35.1±78.1
	All	12.8±6.1	24.8±20.0
DAG [8]	5	16.4±39.5	35.3±53.2
	50	6.2±4.2	21.5±19.8
	All	5.6±3.6	18.4±17.7
SCN [46]	All	6.1±4.0	21.0±21.9
SAM	1	13.1±32.3	32.2±26.2

TABLE III: Results on the lesion matching task.

Method	Lesion emb. [14]	LesaNet [17]	DLT [3]	SAM
Accuracy (%)	80.7	82.1	82.1	91.1

HRNet-5, HRNet-50, and DAG-5 ($p < 0.05$). In pelvic X-ray, SAM is significantly better than HRNet-5 ($p < 10^{-71}$), but not significant compared to HRNet-50 ($p = 0.30$) and DAG-5 ($p = 0.09$). Nevertheless, SAM only used one labeled template compared to DAG using five. Supervised methods are dedicated to modeling specific landmarks, thus may take advantage of the labeled samples better than self-supervised methods. However, they can only detect what has been *a priori* labeled, while SAM can detect arbitrary anatomical locations.

For the lesion matching task, we assess two supervised lesion embedding methods trained on weak labels [14] or labels mined from radiological reports [17]. Since they need the bounding-box of the lesion as input, we use a lesion detector VULD [59] to first detect all lesion candidates in the query CT, and then match them with the template lesion. As shown in Table III, the accuracies of lesion embedding and LesaNet are both significantly lower than SAM. The bottleneck of these detection+matching methods is detection, since it is difficult to find all lesion candidates (especially when they are small and subtle). The detection recall is only 87.3% in our experiment. Therefore, our direct matching strategy is more suitable by avoiding the intermediate detection step. We also compare against a recent lesion tracking algorithm, i.e., Deep Lesion Tracker (DLT) [3]. DLT first initializes the lesion location in the query CT using image registration, then refines the location using a supervised Siamese network. Its accuracy is almost 10% lower than SAM on the 3,556 lesion pairs. Failing to accurately locate some small lesions is the main reason for DLT's errors. DLT was trained on lesions from the official annotations of DeepLesion, whose size distribution is larger than our new test set. This further indicates that supervised methods are limited by their training labels.

For the registration task, we conduct atlas-based segmentation on two datasets: ChestCT (35 annotated organs on 90 test image pairs) and an abdominal CT dataset (13 organs on 45 test pairs) [56]. Average Dice scores and time cost are reported in Table IV. First, SAM-based affine registration significantly

TABLE IV: Comparison of methods on the registration task [18]. For ChestCT, within-phase (CE-CE) and cross-phase (CE-NC) registration are tested. Dice scores (%) are reported.

Method	ChestCT dataset		Abdominal CT dataset	
	CE-CE	CE-NC	Time (s)	
Affine [54]	28.4	28.0	3.4	21.2
FFD [11]	49.4	48.2	93.5	33.4
ANTs (SyN) [9]	49.8	48.0	74.3	28.4
DEEDS [10]	52.7	51.2	45.4	46.5
VoxelMorph [12]	42.7	40.8	4.0	28.5
SAM-affine (proposed) [18]	33.8	33.8	0.5	29.1
SAME (proposed) [18]	54.4	51.0	1.2	49.8
3D U-Net (supervised) [60]	79.7	74.1	109.3	76.8

surpasses traditional optimization-based affine registration in accuracy and speed, suggesting it is a promising registration initialization method. Our proposed SAME (SAM-affine + SAM-coarse + SAM-VoxelMorph) [18] outperforms other methods on most metrics, while only taking 1.2s. It benefits from SAM's fast and accurate point matching results and the semantic information contained in the SAM embeddings. See more details in [18]. **The 3D U-Net [60] is trained on all labeled samples for 500 epochs to converge sufficiently, thus is a strong upper bound. Note that the registration algorithms are all unsupervised. Our proposed SAME is much faster than 3D U-Net since the latter splits the whole volume to overlapped patches to inference separately for better accuracy, which induces more time overhead.**

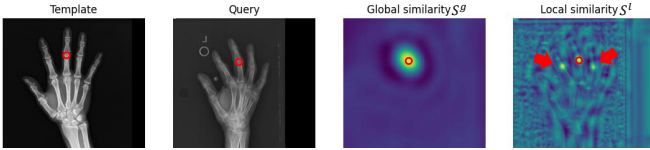
Besides these applications, SAM can potentially be used to aid other medical image analysis tasks. For example, we have proposed a framework to classify CT contrast phases in [61]. We first apply SAM to detect contrast-related landmarks, like the aorta, using template point annotations from only one patient. Then we extract Hounsfield Unit values from the landmarks as features and finally train an SVM for classification. This simple method outperforms a deep learning baseline [62] (F1 95.9% vs. 94.5%) while being much more interpretable.

E. Ablation and Parameter Study

Coarse-to-fine embedding architecture: Table V outlines the ablation study results on the five tasks (landmark detection on ChestCT CE, ChestCT NC, hand X-ray, pelvic X-ray; and universal lesion matching). As can be seen, the coarse-to-fine embedding learning strategy is important to SAM. For instance, we test only learning one fine-scale embedding for each pixel and remove the global embedding, which is similar to other dense self-supervised methods [21], [22]. As shown in row (b), the results degrade in all tasks. In this setting, the fine-scale embedding has to encode both global and local anatomical information, which is more challenging than our coarse-to-fine strategy. We also train both global and local embeddings, but only use one of them for inference. As shown in rows (c) and (d), the results are not promising. The global embedding can provide a rough localization, but it is not locally accurate (see the small offsets between the

TABLE V: Ablation study of the proposed method. The best result in each metric is shown in bold. CE: contrast-enhanced; NC: non-contrast; MRE: mean radial error \pm std.; Max: maximum radial error. Errors are calculated in pixels.

Method	ChestCT CE		ChestCT NC		Hand X-ray	Pelvic X-ray	Universal lesion matching		
	MRE	Max	MRE	Max	MRE	MRE	Accuracy (%)	MRE	Max
(a) SAM (proposed)	4.3\pm3.0	16.4	4.5\pm3.0	18.5	13.1\pm32.3	32.2\pm26.2	91.1	2.7\pm2.5	28.8
(b) w/o coarse-to-fine structure	4.9 \pm 3.0	17.4	5.2 \pm 4.3	61.2	26.1 \pm 70.0	34.4 \pm 31.6	84.1	3.5 \pm 3.1	41.1
(c) Test: Global embedding only	9.9 \pm 12.1	80.1	9.4 \pm 10.4	79.3	437.0 \pm 431.1	48.8 \pm 28.8	52.5	6.6 \pm 3.2	29.9
(d) Test: Local embedding only	6.3 \pm 12.2	150.1	6.5 \pm 9.2	84.7	473.4 \pm 519.2	34.6 \pm 43.3	90.1	3.2 \pm 6.6	130.4
(e) w/o hard negative sampling	4.5 \pm 3.2	17.6	4.7 \pm 3.2	16.6	19.1 \pm 46.3	39.7 \pm 33.0	89.0	2.9 \pm 2.7	26.9
(f) w/o diverse negative sampling	5.3 \pm 6.6	58.7	7.6 \pm 11.6	83.4	14.7 \pm 39.5	45.1 \pm 41.1	92.2	2.7 \pm 3.4	58.3
(g) w/o cutting FPN connection	4.7 \pm 5.3	60.2	5.9 \pm 9.1	79.9	13.9 \pm 22.2	37.2 \pm 28.2	88.6	3.0 \pm 2.9	44.8
(h) w/o ImageNet pretraining	4.6 \pm 3.2	18.7	4.8 \pm 3.2	18.7	17.6 \pm 37.6	38.1 \pm 29.0	88.0	3.0 \pm 2.6	26.8
(i) U-Net backbone [60], [63]	5.9 \pm 5.8	49.3	6.0 \pm 5.6	47.9	37.2 \pm 74.2	41.2 \pm 42.6	84.7	3.5 \pm 3.4	42.2
(j) SAM trained on NIH-LN alone	4.5 \pm 3.3	20.8	4.9 \pm 3.7	24.2	–	–	91.4	2.7 \pm 2.6	30.7

**Fig. 7:** An example in the hand X-ray dataset. Red circles show the ground-truth landmarks. Arrows indicate spurious highlights in the local similarity map of the query image.

global similarity peaks and the red points in Fig. 1). When using it alone, mean accuracies of all tasks drop considerably. The local embedding can localize more precisely, but it may also highlight distant areas with similar local textures if the global similarity is not considered, leading to a large maximum error, which will decrease the algorithm’s stability when it is used in practice. This can be further illustrated in Fig. 7. Neighboring finger joints display similar local texture and may be mismatched by the local embedding, explaining the large error of this dataset in row (d).

Sampling strategies: Removing the hard negative sampling strategy lowers average accuracy in all tasks, as shown in row (e). The difference is more evident in the two X-ray datasets, where the mean errors increase from 13.1 and 32.2 to 19.1 and 39.7, respectively. This is possibly because the appearance of each pixel is less distinctive in X-ray images compared to CT images (Fig. 5), so hard negative pixel sampling is more important for X-rays. Row (f) outlines performance when removing the diverse negative sampling strategy and only sampling hard negative pixels in training. We can find that the mean errors increase in the landmark detection tasks, and the maximum errors increase considerably. This is possibly because the model overfits to the hard examples in the training set, so it produces more distant false highlights in the test set, reducing the generalizability and stability of the algorithm.

Backbone and training strategies: The backbone of SAM is a customized FPN initialized with ImageNet pretrained weights using the inflated 3D technique [41]. The connection between the coarsest FPN feature and its upper level is cut to make global and local embeddings more independent, forcing

TABLE VI: MRE with different data augmentation used in training. We start with random crop and add more augmentations step-by-step. Def. & Rot.: deformation and rotation.

Method	ChestCT CE	ChestCT NC	Hand X-ray	Pelvic X-ray	Lesion match.
Crop only	4.7	4.7	26.7	32.6	2.8
+ Scaling	4.3	4.6	21.9	32.9	2.7
+ Intensity jitter	4.3	4.5	22.6	33.8	2.7
+ Def. & Rot.	4.5	4.6	13.1	32.2	2.6
+ Flip	4.6	5.0	64.9	326.4	2.8

the local embeddings focus on fine-level features. In row (g) of Table V, we do not cut the connection and observe accuracy drop in all datasets. In row (h), the backbone is randomly initialized and trained for twice as many epochs. The accuracy dropped slightly, indicating that ImageNet initialization is helpful. We have also tried other backbone choices such as 3D and 2D U-Nets [60], [63] with a maximum channel number of 512 and a longer training schedule than SAM to compensate for random initialization. However, the accuracy of U-Net is not as good as SAM with random initialization (row (i) vs. (h)), demonstrating the superiority of the proposed backbone. In row (g), we train SAM on the NIH-LN dataset alone and test it on DeepLesion and ChestCT. Its performance only degrades slightly compared to row (a), which uses DeepLesion and ChestCT in training. This result indicates that the embeddings learned in SAM are generalizable across datasets. It is also thanks to the fact that the intrinsic structure of human organs are relatively stable in radiological images, which is the basis of this work.

Data augmentation: Table VI analyzes the effect of different data augmentation approaches, which is key to contrastive learning [6]. Results show that random cropping and scaling suffice to learn good embeddings in CT, while random deformation and rotation (within 30°) is crucial for hand X-ray, as rotated and deformed hands are common in the dataset. Random flipping harms performance as the medical images

we used are normalized in anatomical direction. For pelvic X-rays, it is hard to distinguish their left and right sides, so flipped images confuse model training. Color distortion is found helpful for natural images in [6]. However, intensity jitter brought little gain in our experiments, probably because the intensity distribution is relatively stable in CT and X-ray. We hypothesize that a data augmentation approach for learning effective anatomical embedding is essential only if the corresponding image variance commonly exists in the data.

Parameter study: The influence of parameter values is investigated in Fig. 8. We can see that SAM is robust to embedding dimensions, although larger dimension slightly improves performance [6]. The size of the image patch in z -axis should not be too small (≥ 32 pixels). SAM is robust to n_{pos} and n_{neg} as long as $n_{\text{neg}} \geq 100$. SAM is robust to the overlap ratio of the two augmented patches in training, but the overlap upper bound should not be too small.

F. Comparison with Other Feature Extractors

In this section, we compare SAM with several feature baselines including a traditional handcrafted descriptor, supervised or self-supervised networks pretrained from other tasks, and random initialization. BRIEF [47] is a handcrafted feature descriptor computed using random intensity difference tests. Med3D [48] is a pretrained 3D encoder network for medical images. We adopt a ResNet-18 model pretrained on 23 datasets². Models Genesis [24] is a self-supervised framework that learns features from unlabeled medical images by image restoration. A released 3D U-Net model trained on chest CT data³ is utilized. All the features above are matched using L2 distance. Finally, a SAM model with random weights is also compared as a baseline. The results are exhibited in Table VII. We can see that SAM obtains the best accuracy in all tasks. BRIEF is not suitable for anatomical point matching in radiological images, possibly because it cannot model global information. Furthermore, the local textures in radiological images are not as discriminative as those in natural images, which brings challenges to a simple gradient-based feature descriptors. Both Med3D and Models Genesis are designed for network pretraining on medical images. They may have learned useful radiological features for downstream tasks, but they are not optimal for the task of anatomical point matching, so there is a clear performance gap between them and SAM.

G. Network Initialization from SAM

Network initialization is the purpose of many self-supervised learning works in medical imaging [24], [28], [30]. Although our goal is learning anatomical embeddings, we wonder whether SAM can also be used for network initialization. We run liver and liver tumor segmentation experiments on the LiTS dataset [64] to verify this idea. Among the 131 labeled CTs, we randomly select 105 for training and the rest for testing. Another experiment is done by training on only 10 labeled samples to simulate a small training set. The

3D U-Net [60] with combined cross entropy and Dice loss is used as the segmentation framework. We compare two ways to initialize the U-Net: random, or from a pretrained SAM model. In both ways, we train using RADam [53] for 500 epochs in the experiment with 105 training samples to make the network converge sufficiently, meanwhile in the experiment with 10 samples we only train for 100 epochs to prevent overfitting. The SAM model is trained using the same method as in Sec. IV-B, except that we change the backbone to 3D U-Net.

From Table VIII we can see that initialization from SAM consistently outperforms random initialization, especially when there are fewer training samples. Besides, fine-tuning from SAM benefits liver tumor segmentation more than liver segmentation, which is possibly because liver tumor segmentation is a harder task and requires more fine-scale local features, while the SAM model already learned good fine-scale features from the pixel discrimination task. These results demonstrate that SAM has the potential to be used for network initialization, although it is not designed for the purpose. Comparing SAM with other self-supervised network initialization algorithms is out of the scope of this paper.

H. Incorporating Auxiliary Label Information

Our main goal is to learn universal anatomical embeddings from unlabeled radiological images. Meanwhile, it is interesting to see whether a few labeled samples can improve the landmark detection accuracy of SAM. Thus, we make attempts to extend SAM to the semi-supervised setting. Specifically, when a training sample is unlabeled, the original unsupervised selection strategy will be used to select positive pairs from two augmented image patches. When a training sample contains labeled landmarks, we select the corresponding landmarks from another random labeled image to form a positive pair. In this way, the label information is incorporated, which may help SAM to learn “cross-image” similarity better. We conduct this supervised positive selection strategy for the labeled samples with a probability (0.5 in our experiments), because we find that using supervised selection all the time will cause overfitting and degrade the performance. Results are shown in Table IX. We observe that SAM does benefit from few-shot labeled samples in both 3D CT and 2D X-ray tasks. When 50 labeled samples are used, SAM is comparable with HRNet trained with all labeled samples (Table II). How to leverage label information more effectively will be a future work. Instead of improving SAM with label information, another future direction may be improving landmark detection algorithms with SAM-style self-supervision, such as network initialization (Sec. IV-G) or a regularization loss term.

V. CONCLUSION

In this paper, we propose self-supervised anatomical embedding (SAM) to learn pixel-wise anatomical representation from unlabeled radiological images. It is fast and easy to implement. SAM can be used to locate arbitrary body parts with only one labeled template. The learned embeddings are robust to common image variations, applicable to various image modalities and body parts, and generalizable. It can

²<https://github.com/Tencent/MedicalNet>

³<https://github.com/MrGiovanni/ModelsGenesis/tree/master/pytorch>

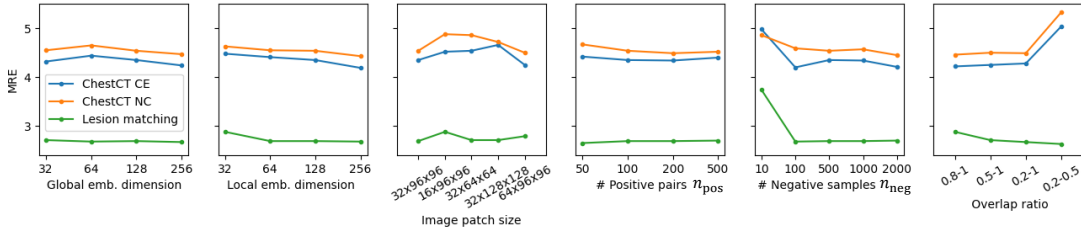


Fig. 8: Parameter study on the ChestCT and lesion matching datasets. Lower MRE is better.

TABLE VII: Comparison of different feature extraction algorithms. “-” means the model is 3D but the task is 2D.

Method	ChestCT CE		ChestCT NC		Hand X-ray	Pelvic X-ray	Universal lesion matching		
	MRE	Max	MRE	Max	MRE	MRE	Accuracy (%)	MRE	Max
(a) BRIEF [47]	23.3±35.2	153.2	19.1±28.5	148.1	324.9±376.3	777.0±618.0	62.4	20.6±33.8	177.0
(b) Med3D [48]	14.0±13.8	67.0	14.4±14.1	63.1	-	-	54.1	16.6±25.1	188.5
(c) Models genesis [24]	8.7±8.1	53.3	7.7±7.4	64.7	-	-	47.5	5.6±7.1	85.2
(d) SAM with random weights	8.7±10.0	91.3	7.8±8.1	78.2	112.0±158.1	320.0±357.2	67.1	7.1±10.2	101.6
(e) SAM (proposed)	4.5±3.2	17.6	4.7±3.2	16.6	19.1±46.3	39.7±33.0	89.0	2.9±2.7	26.9

TABLE VIII: Comparison of different network initialization methods. The liver and liver tumor segmentation accuracy on the LiTS dataset [64] is shown. ASD: average surface distance.

# Labels	Initialization	Liver		Live tumor	
		Dice	ASD	Dice	ASD
105	Random	0.97±0.02	1.4±1.3	0.62±0.13	1.9±2.3
	From SAM	0.97±0.03	1.3±1.4	0.64±0.13	1.8±3.0
10	Random	0.86±0.08	3.2±2.6	0.42±0.34	5.7±4.5
	From SAM	0.90±0.04	2.1±3.8	0.55±0.23	4.6±6.9

TABLE IX: MRE of SAM with label information. $x+1$ means x labeled samples used in training and 1 used as inference template. 20/50 means 20 for ChestCT and 50 for X-ray.

# Labeled samples	ChestCT CE	ChestCT NC	Hand X-ray	Pelvic X-ray
0+1	4.3±3.0	4.5±3.0	13.1±32.3	32.2±26.2
5+1	4.1±2.9	4.2±2.9	11.7±25.7	31.8±24.4
20/50+1	3.9±2.7	4.0±2.8	9.5±12.8	27.9±22.6

be used in universal landmark detection, lesion matching, and registration. Comprehensive experimental results showed that SAM consistently outperforms other registration algorithms and supervised learning methods in few-shot scenarios.

REFERENCES

- [1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, no. 1995, pp. 60–88, 2017.
- [2] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, and D. Comaniciu, “Robust landmark detection in volumetric data with efficient 3D deep learning,” in *MICCAI*, 2015, no. 9783319429984, pp. 49–61.
- [3] J. Cai, Y. Tang, K. Yan, A. P. Harrison, J. Xiao, G. Lin, and L. Lu, “Deep Lesion Tracker: Monitoring Lesions in 4D Longitudinal Imaging Studies,” in *CVPR*, 2021.
- [4] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. Chiang, Z. Wu, and X. Ding, “Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation,” *Medical Image Analysis*, 2020.
- [5] Z. Wu, Y. Xiong, S. Yu, and D. Lin, “Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination,” in *CVPR*, 2018.
- [6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” in *ICML*, 2020.
- [7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum Contrast for Unsupervised Visual Representation Learning,” in *CVPR*, 2020, pp. 9726–9735.
- [8] W. Li, Y. Lu, K. Zheng, H. Liao, C. Lin, J. Luo, C.-T. Cheng, J. Xiao, L. Lu, C.-F. Kuo, and S. Miao, “Structured Landmark Detection via Topology-Adapting Deep Graph Learning,” in *ECCV*, 2020.
- [9] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, “Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain,” *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, 2008. [Online]. Available: www.itk.org
- [10] M. P. Heinrich, M. Jenkinson, M. Brady, and J. A. Schnabel, “Mrf-based deformable registration and ventilation estimation of lung ct,” *IEEE Trans. Med. Imaging*, vol. 32, no. 7, pp. 1239–1248, 2013.
- [11] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, “Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images,” *IEEE Trans. Med. Imaging*, vol. 18, no. 8, 1999.
- [12] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, “VoxelMorph: A Learning Framework for Deformable Medical Image Registration,” *IEEE Trans. Med. Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.
- [13] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep High-Resolution Representation Learning for Human Pose Estimation,” in *CVPR*, 2019.
- [14] K. Yan et al., “Deep Lesion Graphs in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-scale Lesion Database,” in *CVPR*, 2018.
- [15] X. Rafael-Palou, A. Aubanell, I. Bonavita, M. Ceresa, G. Piella, V. Ribas, and M. A. González Ballester, “Re-Identification and growth detection of pulmonary nodules without image registration using 3D siamese neural networks,” *Med. Image Anal.*, vol. 67, p. 101823, 2021.
- [16] K. Yan, X. Wang, L. Lu, and R. M. Summers, “DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning,” *Journal of Medical Imaging*, vol. 5, no. 3, 2018.
- [17] K. Yan, Y. Peng, V. Sandfort, M. Bagheri, Z. Lu, and R. M. Summers, “Holistic and comprehensive annotation of clinically significant findings on diverse CT images: Learning from radiology reports and label ontology,” in *CVPR*, vol. 2019-June, 2019, pp. 8515–8524.
- [18] F. Liu, K. Yan, A. P. Harrison, D. Guo, L. Lu, A. Yuille, L. Huang,

- G. Xie, J. Xiao, X. Ye, and D. Jin, "SAME: Deformable Image Registration based on Self-supervised Anatomical Embeddings," in *MICCAI*, 2021 (attached as a supplementary file).
- [19] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *ICLR*, 2018.
- [20] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," in *CVPR*, vol. 2016-Decem, 2016, pp. 2536–2544.
- [21] P. O. Pinheiro, A. Almahairi, R. Y. Benmaleck, F. Golemo, and A. Courville, "Unsupervised Learning of Dense Visual Representations," in *NeurIPS*, 2020.
- [22] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *CVPR*, 2021.
- [23] A. Van Den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018.
- [24] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models Genesis," *Med. Image Anal.*, 2020.
- [25] R. Feng, Z. Zhou, M. B. Gotway, and J. Liang, "Parts2Whole: Self-supervised Contrastive Learning via Reconstruction," in *MICCAI Work. Domain Adapt. Represent. Transf.*, 2020.
- [26] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Med. Image Anal.*, vol. 58, p. 101539, 2019.
- [27] M. Blendsowski, H. Nickisch, and M. P. Heinrich, "How to Learn from Unlabeled Volume Data: Self-supervised 3D Context Feature Learning," in *MICCAI*, vol. 11769 LNCS, 2019, pp. 649–657.
- [28] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, "Self-supervised feature learning for 3d medical images by playing a rubik's cube," in *MICCAI*, vol. 11767 LNCS, 2019, pp. 420–428.
- [29] N. Tajbakhsh, Y. Hu, J. Cao, X. Yan, Y. Xiao, Y. Lu, J. Liang, D. Terzopoulos, and X. Ding, "Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data," in *ISBI*, vol. 2019-April, 2019, pp. 1251–1255.
- [30] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi, "Big Self-Supervised Models Advance Medical Image Classifications," in *ICCV*, 2021.
- [31] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *NeurIPS*, 2020.
- [32] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, "Self-supervision with Superpixels: Training Few-shot Medical Image Segmentation without Annotation," in *ECCV*, 2020.
- [33] H. Chen, Y. Wang, K. Zheng, W. Li, C.-T. Cheng, A. P. Harrison, J. Xiao, G. D. Hager, L. Lu, C.-H. Liao, and S. Miao, "Anatomy-Aware Siamese Network: Exploiting Semantic Asymmetry for Accurate Pelvic Fracture Detection in X-ray Images," in *ECCV*, 2020.
- [34] H. R. Roth, C. T. Lee, H.-C. Shin, A. Seff, L. Kim, J. Yao, L. Lu, and R. M. Summers, "Anatomy-specific classification of medical images using deep convolutional nets," pp. 2–5, 2015.
- [35] K. Yan, L. Lu, and R. M. Summers, "Unsupervised Body Part Regression via Spatially Self-ordering Convolutional Neural Networks," in *ISBI*, 2018.
- [36] H. Sokooti, B. de Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring, "Nonrigid image registration using multi-scale 3D convolutional neural networks," in *MICCAI*, vol. 10433 LNCS, 2017, pp. 232–239.
- [37] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *NIPS*, 2015, pp. 2017–2025.
- [38] F. Liu, J. Cai, Y. Huo, C.-T. Cheng, A. Raju, D. Jin, J. Xiao, A. Yuille, L. Lu, C. Liao, and A. P. Harrison, "JSSR: A Joint Synthesis, Segmentation, and Registration System for 3D Multi-Modal Image Alignment of Large-scale Pathological CT Scans," in *ECCV*, 2020.
- [39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016, pp. 770–778.
- [40] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.
- [41] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *CVPR*, vol. 2017-Janua, 2017, pp. 4724–4733.
- [42] M. Blendsowski, H. Nickisch, and M. P. Heinrich, "How to Learn from Unlabeled Volume Data: Self-supervised 3D Context Feature Learning," in *MICCAI*, vol. 11769 LNCS, 2019, pp. 649–657.
- [43] D. Novotny, S. Albanie, D. Larlus, and A. Vedaldi, "Self-Supervised Learning of Geometrically Stable Features Through Probabilistic Introspection," in *CVPR*, 2018, pp. 3637–3645.
- [44] S. Wang, S. Cao, D. Wei, R. Wang, K. Ma, L. Wang, D. Meng, and Y. Zheng, "LT-Net: Label Transfer by Learning Reversible Voxel-wise Correspondence for One-shot Medical Image Segmentation," in *CVPR*, 2020.
- [45] A. Dosovitskiy, P. Fischery, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *ICCV*, vol. 2015 Inter, 2015, pp. 2758–2766.
- [46] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Integrating spatial configuration into heatmap regression based CNNs for landmark localization," *Med. Image Anal.*, vol. 54, pp. 207–219, 2019. [Online]. Available: <https://doi.org/10.1016/j.media.2019.03.007>
- [47] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in *ECCV*, 2010.
- [48] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer Learning for 3D Medical Image Analysis," 2019. [Online]. Available: <http://arxiv.org/abs/1904.00625>
- [49] "CT Lymph Nodes dataset - The Cancer Imaging Archive (TCIA) Public Access," 2016. [Online]. Available: <https://wiki.cancerimagingarchive.net/display/Public/CT+Lymph+Nodes>
- [50] S. Wang, S. Cao, D. Wei, R. Wang, K. Ma, L. Wang, D. Meng, and Y. Zheng, "LT-Net: Label Transfer by Learning Reversible Voxel-wise Correspondence for One-shot Medical Image Segmentation," in *CVPR*, 2020.
- [51] I. Fabian, J. Paul, and W. Jakob, "batchgenerators - a python framework for data augmentation," Tech. Rep., 2020. [Online]. Available: <https://github.com/MIC-DKFZ/batchgenerators>
- [52] F. Pérez-García, R. Sparks, and S. Ourselin, "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," Tech. Rep., 2020. [Online]. Available: <https://github.com/fepegar/torchio>
- [53] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.
- [54] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imaging*, vol. 29, no. 1, pp. 196–205, 2010. [Online]. Available: <http://elastix.isi.uu.nl/wiki.php>
- [55] K. Murphy, B. Van Ginneken, and J. M. Reinhardt *et al.*, "Evaluation of registration methods on thoracic ct: the empire10 challenge," *IEEE Trans. Med. Imaging*, vol. 30, no. 11, pp. 1901–1920, 2011.
- [56] Z. Xu, C. P. Lee, and M. P. Heinrich *et al.*, "Evaluation of six registration methods for the human abdomen on clinically acquired ct," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1563–1572, 2016.
- [57] M. P. Heinrich and L. Hansen, "Highly accurate and memory efficient unsupervised learning-based discrete CT registration using 2.5 D displacement search," in *MICCAI*, 2020.
- [58] X. Hu, M. Kang, W. Huang, M. R. Scott, R. Wiest, and M. Reyes, "Dual-stream pyramid registration network," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds. Springer International Publishing, 2019.
- [59] J. Cai, K. Yan, C.-T. Cheng, J. Xiao, C.-H. Liao, L. Lu, and A. P. Harrison, "Deep Volumetric Universal Lesion Detection Using Light-Weight Pseudo 3D Convolution and Surface Point Regression," in *MICCAI*, 2020, pp. 3–13.
- [60] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *MICCAI*, vol. 9901 LNCS, 2016, pp. 424–432. [Online]. Available: <http://imb.informatik.uni-freiburg.de/resources/opensource/unet.en.html>
- [61] K. Yan, Y. Tang, A. P. Harrison, J. Cai, L. Lu, and J. Lu, "Interpretable Medical Image Classification with Self-Supervised Anatomical Embedding and Prior Knowledge," in *Int. Conf. Med. Imaging with Deep Learn. (MIDL)*, vol. 1, 2021, pp. 2018–2020. [Online]. Available: <https://arxiv.org/abs/2012.02383>
- [62] B. Zhou, A. P. Harrison, J. Yao, C. T. Cheng, J. Xiao, C. H. Liao, and L. Lu, "CT data curation for liver patients: Phase recognition in dynamic contrast-enhanced CT," in *MICCAI - Med. Image Learn. with Less Labels Imperfect Data Work.*, vol. 11795, 2019, pp. 139–147.
- [63] N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *MICCAI*, vol. 9351, pp. 234–241, 2015.
- [64] P. Bilic, P. F. Christ, and E. Vorontsov *et al.*, "The Liver Tumor Segmentation Benchmark (LiTS)," Tech. Rep., 2019. [Online]. Available: <http://arxiv.org/abs/1901.04056>