Multi-scale Self-Supervised Learning for Longitudinal Lesion Tracking with Optional Supervision

Anamaria Vizitiu[1(✉)], Antonia T. Mohaiu[1], Ioan M. Popdan[1], Abishek Balachandran[2], Florin C. Ghesu[3], and Dorin Comaniciu[3]

[1] Advanta, Siemens SRL, Brasov, Romania
anamaria.vizitiu@siemens.com
[2] Digital Technology and Innovation, Siemens Healthineers, Bangalore, India
[3] Digital Technology and Innovation, Siemens Healthineers, Princeton, NJ, USA

Abstract. Longitudinal lesion or tumor tracking is an essential task in different clinical workflows, including treatment monitoring with follow-up imaging or planning of re-treatments for radiation therapy. Accurately establishing correspondence between lesions at different timepoints, recognizing new lesions or lesions that have disappeared is a tedious task that only grows in complexity as the number of lesions or timepoints increase. To address this task, we propose a generic approach based on multi-scale self-supervised learning. The multi-scale approach allows the efficient and robust learning of a similarity map between multi-timepoint image acquisitions to derive correspondence, while the self-supervised learning formulation enables the generic application to different types of lesions and image modalities. In addition, we impose optional supervision during training by leveraging tens of anatomical landmarks that can be extracted automatically. We train our approach at large scale with more than 50,000 computed tomography (CT) scans and validate it on two different applications: 1) Tracking of generic lesions based on the DeepLesion dataset, including liver tumors, lung nodules, enlarged lymph-nodes, for which we report highest matching accuracy of 92%, with localization accuracy that is nearly 10% higher than the state-of-the-art; and 2) Tracking of lung nodules based on the NLST dataset for which we achieve similarly high performance. In addition, we include an error analysis based on expert radiologist feedback, and discuss next steps as we plan to scale our system across more applications.

Keywords: Self-supervised learning · Multi-scale · Longitudinal lesion tracking

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43907-0_55.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 573–582, 2023.
https://doi.org/10.1007/978-3-031-43907-0_55

1 Introduction

Longitudinal lesion or tumor tracking is a fundamental task in treatment monitoring workflows, and for planning of re-treatments in radiation therapy. Based on longitudinal imaging for a given patient it requires establishing which lesions are corresponding (i.e., same lesion, observed at different timepoints), which lesions have disappeared and which are new compared to prior scanning. This information can be leveraged to assess treatment response, e.g., by analyzing the evolution of size and morphology for a given tumor [1], but also for adaptation of (re-)treatment radiotherapy plans that take into account new tumors.

In practice, the development of automatic and reliable lesion tracking solutions is hindered by the complexity of the data (over different modalities), the absence of large, annotated datasets, and the difficulties associated with lesion identification (i.e., varying sizes, poses, shapes, and sparsely distributed locations). In this work, we present a multi-scale self-supervised learning solution for lesion tracking in longitudinal studies using the capabilities of contrastive learning [9]. Inspired by the pixel-wise contrastive learning strategy introduced in [5], we choose to learn pixel-wise feature representations that embed consistent anatomical information from unlabeled (i.e., without lesion-related annotations) and unpaired (i.e., without the use of longitudinal scans) data, overcoming barriers to data collection. To increase the system robustness and emulate the clinician's reading strategies, we propose to use multi-scale embeddings to enable the system to progressively refine the fine-grained location. In addition, as imaging offers contextual information about the human body that is naturally consistent, we design the model to benefit from biologically-meaningful points (i.e., anatomical landmarks). The reasoning behind this strategy is that simple data augmentation methods cannot faithfully model inter-subject variability or possible organ deformations. Hence, we ensure the spatial coherence of the tracked lesion location using well-defined anatomical landmarks.

Our proposed method brings two elements of novelty from a technical point of view: (1) the multi-scale approach for the anatomical embedding learning and (2) a positive sampling approach that incorporates anatomically significant landmarks across different subjects. With these two strategies, the goal is to ensure a high degree of robustness in the computation of the lesion matching across different lesion sizes and varying anatomies. Furthermore, a significant focus and contribution of our research is the experimental study at a very large scale: we (1) train a pixel-wise self-supervised system using a very large and diverse dataset of 52,487 CT volumes and (2) evaluate on two publicly available datasets. Notably, one of the datasets, NLST, presents challenging cases with 68% of lesions being very small (i.e., radius < 5 mm).

2 Background and Motivation

The problem of lesion tracking in longitudinal data is typically divided into two steps: (1) detection of lesions and (2) tracking the same lesion over multiple

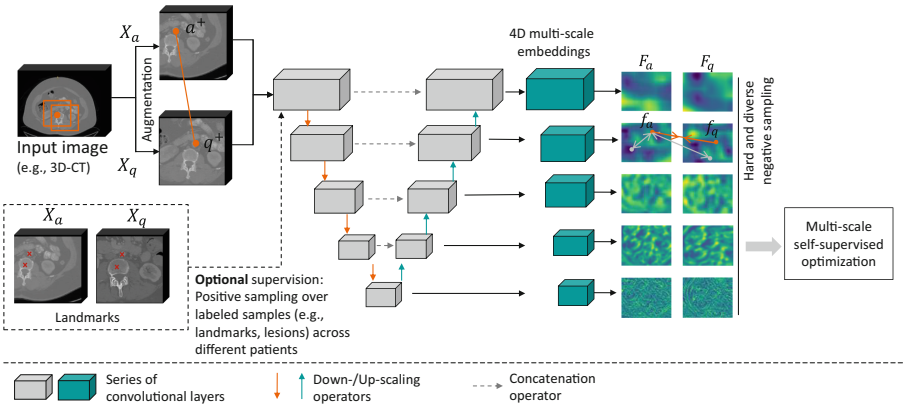

Fig. 1. Overview of the proposed multi-scale self-supervised learning system. During training, we randomly extract positive samples (optionally, include same anatomical landmarks from different volumes), hard-negative samples, and diverse negative samples of pixels from augmented 3D paired patches. During inference, the extracted embeddings are used to generate a cascade of cosine similarity maps that initially locate the corresponding location in a follow-up image within a larger area and subsequently improve the matching accuracy through gradual refinement.

time points. Classical methods to solve this problem rely on image registration, where tracking is performed via image alignment and rule-based correspondence matching [15,16,21]. These approaches are difficult to optimize, especially when scaling across different body regions and fields of view. Appearance-based trackers [19,20] adopt a different strategy by projecting lesions detected beforehand with dedicated detectors [17,18] onto a representation space and employing nearest neighbor analysis. One recent approach, Deep Lesion Tracker (DLT) [8], integrates both strategies to perform appearance-based recognition under anatomical constraints. As a more direct matching approach, Yan et al. [5] uses a self-supervised anatomical embedding model (SAM) to create semantic embeddings for each image pixel, avoiding the detection step. Training exclusively on augmented paired data prevents SAM from accurately representing anatomical changes and deformations that occur over time. This can influence the contextual information of a pixel, which in turn impacts the pixel-wise embeddings on which the similarity-based tracker depends. To overcome this, we propose to train a pixel-wise multi-scale embedding model that accounts for anatomical similarity among different subjects, making the embeddings more effective.

3 Method

3.1 Problem Definition

Let I_1 (i.e., template or baseline image) and I_2 (i.e., query or follow-up image) be two 3D-CT scans acquired at time t_1 and t_2, respectively, Additionally, let p_1

and p_2 denote the point of interest (i.e., the lesion center) in both images. The problem of lesion tracking can be formulated as finding the optimal transformation that maps p_1 to its corresponding location, p_2, in I_2.

3.2 Training Stage

Let $D = \{X_1, X_2, ..., X_N\}$ be a set of N unpaired and unlabeled 3D-CT volumes. As shown in Fig. 1, given an image $X \in \mathbb{R}^{d \times h \times w}$ from the training dataset D, we randomly select two overlapping 3D patches (anchor and query), namely X_a and X_q. To create synthetic paired data that mimics appearance changes across different images, we apply random data augmentation (i.e., random spatial and intensity-related transformations) to the content of X_a and X_q. We implement a similar augmentation strategy to that described in [5]. Given X_a and X_q, we use an embedding extraction model to construct a hierarchy of multi-scale semantic embeddings for each image pixel, labeled F_a and F_q respectively. The embedding at ith scale, $1 \leq i \leq s$, is denoted as F_a^i and F_q^i and is represented as a 4D feature map, with an embedding vector of length L associated with each pixel.

Given the nature of contrastive learning, the sampling strategy (extracting negative and positive pixel pairs from augmented 3D paired patches) is essential to achieving discriminative pixel-wise embeddings. We arbitrary sample n_{pos} positive pixel pairs from the overlapping area of X_a and X_q, denoted by $a^+ = \{a_1^+, ..., a_j^+ ..., a_{n_{pos}}^+\}$, $q^+ = \{q_1^+, ..., q_j^+, ..., q_{n_{pos}}^+\}$, $1 \leq j \leq n_{pos}$. To further enhance embeddings, 10% of the positive pixel pairs are derived from biologically-meaningful points across different volumes in the batch. We use data-driven models [14] to extract 37 anatomical landmarks, such as the top right lung, suprasternal notch, tracheal bifurcation, etc. Similar to [5], for each positive pixel pair (a_j^+, q_j^+), we select n_{neg} hard and diverse negative pixels, denoted by $h^- = \{h_1^-, ..., h_k^-, ..., h_{n_{neg}}^-\}$, $1 \leq k \leq n_{neg}$.

Next, at each scale, we extract the embedding vectors for positive and negative pixel pairs from F_a^i, F_q^i, guided by the corresponding locations, a^+, q^+, h^-, which are downsampled to match the scale. We denote the positive embeddings at ith scale at pixel location a_j^+, q_j^+ as $f_{a_j}^i$, $f_{q_j}^i \in \mathbb{R}^L$. Similarly, we denote the negative embeddings at pixel location h_k^- associated to a positive positive pixel pair (a_j^+, q_j^+) as $f_{jk}^i \in \mathbb{R}^L$. We use L2-norm to normalize the embedding vectors before the loss computation. We use pixel-wise InfoNCE loss [5,10] to enhance the similarity among similar pixels (i.e., positive pairs of pixels) and decrease the similarity among dissimilar pixels (i.e., negative pairs of pixels). Correspondingly, we set the contrastive loss at the ith scale:

$$L^i = -\sum_{j=1}^{n_{pos}} \log \frac{\exp(f_{a_j}^i \cdot f_{q_j}^i / \tau)}{\exp(f_{a_j}^i \cdot f_{q_j}^i / \tau) + \sum_{k=1}^{n_{neg}} \exp(f_{a_j}^i \cdot f_{jk}^i / \tau)}, \tag{1}$$

where $\tau = 0.5$ is a temperature parameter. The final loss is then calculated as the average of all these individual losses.

3.3 Inference Stage

Let X_a be a 3D-CT volume template with an input point of interest $p_a \in X_a$, and X_q a corresponding query 3D-CT volume. The first step is to project the image X_a into a multi-scale feature space, creating a hierarchy of multi-scale semantic embeddings F_a for each pixel in the image (i.e., a 4D feature map). Next, we follow a similar process for the query image X_a and acquire the pixel-level embeddings F_q.

To measure the similarity between the embeddings of the input X_a at the point of interest p_a and the query embeddings F_q, we compute cosine similarity maps at each scale:

$$S^i = \frac{F_a^i(p_a) \cdot F_q^i}{\|F_a^i(p_a)\|_2 \cdot \|F_q^i\|_2} \tag{2}$$

Finally, we combine the multi-scale similarity maps through summation and select the voxel with the highest similarity as the matching point in the query volume.

4 Experiments

4.1 Datasets and Setup

Datasets: We train the universal and fine-grained anatomical point matching model using an in-house CT dataset (VariousCT). The training dataset contains 52,487 unlabeled 3D CT volumes capturing various anatomies, including chest, head, abdomen, pelvis, and more.

The evaluation is based on two datasets, the publicly released Deep Longitudinal Study (DLS) dataset [8] and the National Lung Screening Trial (NLST) dataset [12]. The DLS dataset is a subset of the DeepLesion [11] medical imaging dataset, containing 3891 pairs of lesions with information on their location and size. The dataset covers various types of lesions across different organs. We follow the official data split for DLS dataset and perform evaluation on the testing dataset which comprises 480 lesion pairs. For NLST, we randomly selected a subset of 1045 test images coming from 420 patients with up to 3 studies. A certified radiologist annotated the testing data by identifying the location and size of the pulmonary nodules, resulting in a total of 825 paired annotations. We evaluate lesion tracking in both directions, from baseline to follow-up and from follow-up to baseline [8]. This results in a total of 960 and 1650 testing lesion pairs in DLS and NLST test sets, respectively. The isotropic resolution of all CT volumes is adjusted to $2mm$ through bilinear interpolation.

System Training: Our learning model is implemented in PyTorch and uses the TorchIO library [13] for medical data manipulation and augmentation.

We employ a U-Net-based encoder-decoder architecture [2] that utilizes an inflated 3D ResNet-18 [3,4] as its encoder, which extends all 2D convolutions

Table 1. Comparison between the proposed solution and several state-of-the-art approaches (reference results are from [8]). The exact same test set was used to compute the performance of each approach listed in the table; however, we retrained only SAM.

Method	CPM@ 10 mm	CPM@ Radius	MEDx (mm)	MEDy (mm)	MEDz (mm)	MED (mm)
Affine [15]	48.33	65.21	4.1 ± 5.0	5.4 ± 5.6	7.1 ± 8.3	11.2 ± 9.9
VoxelMorph [16]	49.90	65.59	4.6 ± 6.7	5.2 ± 7.9	6.6 ± 6.2	10.9 ± 10.9
LENS-LesionG. [17,19]	63.85	80.42	2.6 ± 4.6	2.7 ± 4.5	6.0 ± 8.6	8.0 ± 10.1
VULD-LesionG. [18,19]	64.69	76.56	3.5 ± 5.2	4.1 ± 5.8	6.1 ± 8.8	9.3 ± 10.9
LENS-LesaNet [17,20]	70.00	84.58	2.7 ± 4.8	2.6 ± 4.7	5.7 ± 8.6	7.8 ± 10.3
DLT-SSL [8]	71.04	81.52	3.8 ± 5.3	3.7 ± 5.5	5.4 ± 8.4	8.8 ± 10.5
DEEDS [21]	71.88	85.52	2.8 ± 3.7	3.1 ± 4.1	5.0 ± 6.8	7.4 ± 8.1
DLT-Mix [8]	78.65	88.75	3.1 ± 4.4	3.1 ± 4.5	4.2 ± 7.6	7.1 ± 9.2
DLT [8]	78.85	86.88	3.5 ± 5.6	2.9 ± 4.9	4.0 ± 6.1	7.0 ± 8.9
SAM [5]	81.67	90.21	2.9 ± 8.0	2.5 ± 3.8	3.6 ± 5.2	6.5 ± 9.6
Ours	**83.13**†	**91.87**†	$\mathbf{2.9 \pm 6.0}$	$\mathbf{2.2 \pm 3.2}$	$\mathbf{3.1 \pm 3.9}$	$\mathbf{5.9 \pm 7.1}^{\dagger}$

† Improvement is statistically significant compared to SAM [5] (p-value $< 10^{-6}$).

in the standard ResNet to 3D convolutions and allows the use of pre-trained ImageNet weights. The multi-scale embedding model employs $s = 5$ scales, and the embedding length is fixed at $L = 128$ for each scale. Convolution with a stride of $(2, 2, 2)$ is used to reduce the feature map size at the first and fifth levels, while a stride of $(1, 2, 2)$ is employed for intermediary levels 2 to 4. The U-Net decoder uses a convolution layer with a $3 \times 3 \times 3$ kernel after every up-sampling layer to generate the final cascade of feature embeddings.

The model is trained with AdamW optimizer [6] for 64 epochs using an early stopping strategy with a patience of 5 epochs, a batch size of 8 augmented 3D paired patches of $32 \times 96 \times 96$, and a learning rate of 0.0001.

For data augmentation, we apply random cropping, scaling, rotation, and Gaussian noise injections. A windowing approach that covers the intensity ranges of lungs and soft tissues is used to scale CT intensity values to $[-1, 1]$. The sampling hyperparameters consist of 100 positive pixel pairs ($n_{pos} = 100$), 100 hard negative pixel pairs, and 200 diverse negative pixel pairs ($n_{neg} = 300$).

Evaluation Metrics: We use mean Euclidean distance (MED) to measure the distance between predicted lesion center and ground truth, and the center point matching accuracy (i.e., percentage of accurately matched lesions given the annotated lesion radius), denoted with CPM@Radius. For lesions of large sizes, we set a maximum distance limit of 10 mm as acceptance criteria [8], denoted with CPM@10 mm. The NLST testing dataset has a distinctive feature wherein nodules are relatively small, 68% of annotated lesions have a radius of less than 5 mm (compared to 6% in DLS dataset). To ensure that such small nodules are not missed during evaluation, we relax the minimum distance requirement and consider a distance of 6 mm as a permissible matching error.

4.2 Evaluation

For the lesion tracking task on DLS dataset, we quantitatively compare our system against existing trackers in Table 1. These include the Deep Lesion Tracker (DLT) and its variants [8], as well as registration-based trackers [15,16,21] and appearance-based trackers via detector learning [17–20]. Given the clear superiority of approach [5] compared to all reference solutions, we focus on achieving a direct comparison against SAM [5]. Hence, for performance comparison against self-supervised anatomical embedding tracker, we retrain SAM [5] with images from VariousCT dataset.

Table 2. Results on the NLST dataset related to the tracking of lung nodules.

Method	CPM@ 10 mm	CMP@ Radius	MEDx (mm)	MEDy (mm)	MEDz (mm)	MED(mm)
Ours	**90.05**	**92.12**	$\mathbf{2.0 \pm 2.6}$	$\mathbf{2.2 \pm 3.8}$	$\mathbf{2.7 \pm 4.8}$	$\mathbf{4.9 \pm 6.0}$

Our method achieves a matching accuracy of 91.87%, that is 1.84% higher than SAM and 5.74% higher than DLT. To confirm the significance of the improvement achieved by our method compared to SAM [5], we conduct a paired t-test for statistical analysis and show that the improvement is statistically significant (p-value $< 10^{-6}$). Compared to the self-supervised version of DLT, the difference in performance is significantly greater, the proposed systems outperforms DLT-SSL by more than 10%. When imposing a maximum distance limit of 10 mm between the ground truth and prediction, our method increases performance by 1.46%, showing the importance of the multi-scale approach in lesion

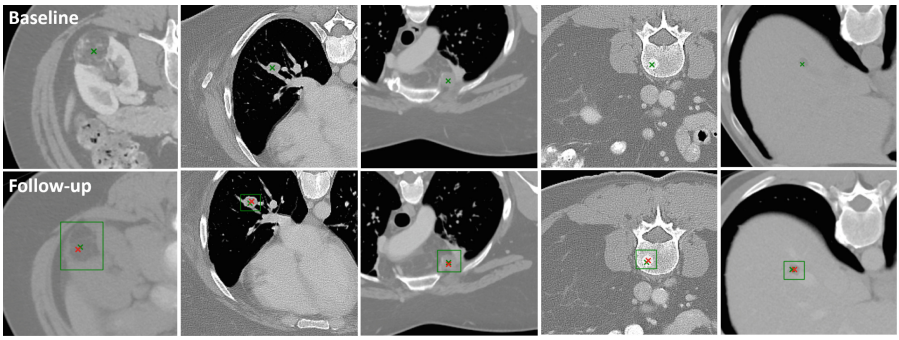

Fig. 2. Examples of lesion matching results on the DLS testing dataset. We denote the ground-truth points using green markers in both the baseline and follow-up images, whereas the predicted points are indicated by red markers. To illustrate the extent of the lesions, we also display the annotated bounding boxes on the follow-up images. For more clarity, we show only the axial view. (Color figure online)

location refinement. Consequently, superior accuracy is achieved compared to both registration and supervised appearance-based tracking methods. Qualitative examples are shown in Fig. 2.

On the NLST dataset, our proposed method obtains a center point matching accuracy of 92.12% (Table 2). In the case of longitudinal lung nodule tracking (Fig. 3), it is more frequent to observe significant changes in size and density. As our system relies on the concept of anatomical embedding matching, the most substantial errors in lesion matching for our system occur when there are significant pathological distortions that deviate greatly from one timepoint to another. Examples of such cases are depicted in Fig. 4, based on expert radiologist feedback.

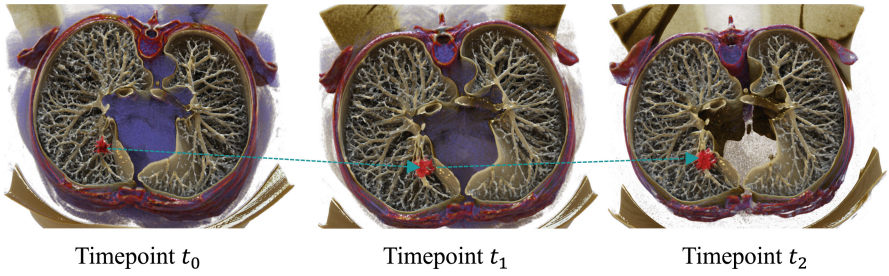

Fig. 3. Example case from the test set, highlighting the progression of a lung nodule over three timepoints: t_0, t_1 and t_3. Our system was robust to the axial rotations of the scans and the increasing size of the nodule and correctly established the correspondence.

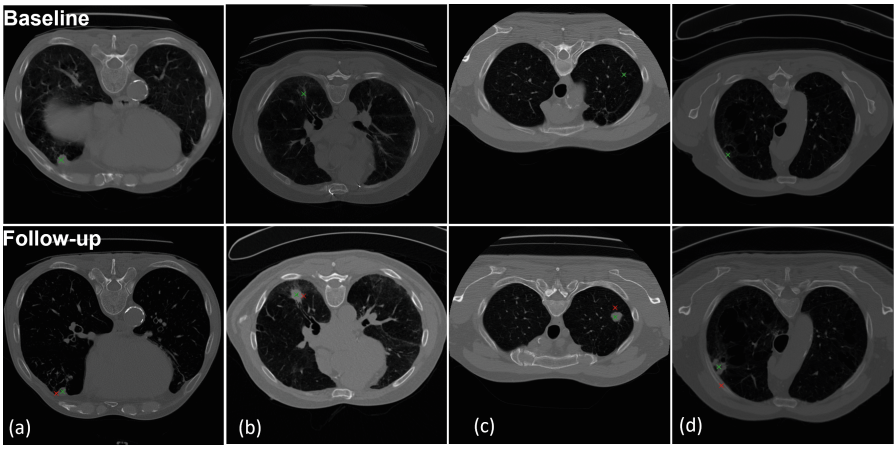

Fig. 4. Examples of lesion matching results on clinically challenging cases from NLST testing dataset: (a) bronchiectasis with mucus plugging adjacent to the nodule, (b) spiculated nodule in a setting of interstitial lung disease, (c), (d) small nodule progressed and increased significantly in size. The green and red markers denote the ground-truth and predicted lesion location. (Color figure online)

5 Conclusion

In conclusion, this paper presents an effective method for longitudinal lesion tracking based on multi-scale self-supervised learning. The method is generic, it does not require expert annotations or longitudinal data for training and can generalize to different types of tumors/organs/modalities. The multi-scale approach ensures a high degree of robustness and accuracy for small lesions. Through large-scale experiments and validation on two longitudinal datasets, we highlight the superiority of the proposed method in comparison to state-of-the-art. We found that adopting a multi-scale approach (instead of the global/local approach as proposed in [5]) can lead to embeddings that better capture the anatomical location and are able to handle lesions that vary in size or appearance at different scales. Moreover, the changes proposed in this work help to alleviate the confusion caused by left-right body symmetries (e.g., the apices of the lungs). This effect challenged the tracking of small nodules in the lungs using [5]. Our future work aims to enhance the matching accuracy by examining the implications of correlation magnitude, conducting robustness studies on slight variations in tracking initialization, and implementing a more advanced fusion strategy for the multi-scale similarity maps. In addition, we aim to expand to more applications, e.g., treatment monitoring for brain cancer using MRI.

Disclaimer: The concepts and information presented in this paper/presentation are based on research results that are not commercially available. Future commercial availability cannot be guaranteed.

Acknowledgements. The authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial (NLST). The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

References

1. Eisenhauer, E.A., et al.: New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur. J. Cancer **45**(2), 228–247 (2009)
2. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR, pp. 4724–4733 (2017)
5. Yan, K., et al.: SAM: self-supervised learning of pixel-wise anatomical embeddings in radiological images. IEEE Trans. Med. Imaging **41**, 2658–2669 (2020)
6. Ilya, L., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017)

7. Ilya, L., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. ArXiv: Learning (2016)
8. Cai, J., et al.: Deep lesion tracker: monitoring lesions in 4D longitudinal imaging studies. In: CVPR, pp. 15154–15164 (2020)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations (2020)
10. Van Den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2018)
11. Yan, K., Wang, X., Lu, L., Summers, R.M.: DeepLesion: automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations (2017)
12. National Lung Screening Trial Research Team: The National Lung Screening Trial: overview and study design. In: Radiology, pp. 243–253 (2011)
13. Perez-Garcia, F., Sparks, R., Ourselin, S.: TorchIO: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput. Methods Prog. Biomed. **208**, 106236 (2020)
14. Ghesu, F.C., et al.: Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 176–189 (2019)
15. Marstal, K., Berendsen, F.F., Staring, M., Klein, S.: SimpleElastix: a user-friendly, multi-lingual library for medical image registration. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 574–582 (2016)
16. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J.V., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: CVPR, pp. 9252–9260 (2018)
17. Yan, K., et al.: Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in CT. IEEE Trans. Med. Imaging **40**, 2759–2770 (2020)
18. Cai, J., et al.: Deep volumetric universal lesion detection using light-weight pseudo 3D convolution and surface point regression (2020)
19. Yan, K., 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: Conference on Computer Vision and Pattern Recognition, pp. 9261–9270 (2017)
20. Yan, K., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., Summers, R.M.: Holistic and comprehensive annotation of clinically significant findings on diverse CT images: learning from radiology reports and label ontology. In: CVPR, pp. 8515–8524 (2019)
21. Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A.: MRF-based deformable registration and ventilation estimation of lung CT. IEEE Trans. Med. Imaging **32**, 1239–1248 (2013)