Transformer Lesion Tracker

Wen Tang[1], Han Kang[1], Haoyue Zhang[1,2], Pengxin Yu[1], Corey W. Arnold[2], and Rongguo Zhang[1(✉)]

[1] InferVision Medical Technology Co., Ltd., Beijing, China
`zrongguo@infervision.com`
[2] Computational Diagnostic Lab, UCLA, Los Angeles, USA

Abstract. Evaluating lesion progression and treatment response via longitudinal lesion tracking plays a critical role in clinical practice. Automated approaches for this task are motivated by prohibitive labor costs and time consumption when lesion matching is done manually. Previous methods typically lack the integration of local and global information. In this work, we propose a transformer-based approach, termed Transformer Lesion Tracker (TLT). Specifically, we design a Cross Attention-based Transformer (CAT) to capture and combine both global and local information to enhance feature extraction. We also develop a Registration-based Anatomical Attention Module (RAAM) to introduce anatomical information to CAT so that it can focus on useful feature knowledge. A Sparse Selection Strategy (SSS) is presented for selecting features and reducing memory footprint in Transformer training. In addition, we use a global regression to further improve model performance. We conduct experiments on a public dataset to show the superiority of our method and find that our model performance has improved the average Euclidean center error by at least 14.3% (6 mm vs. 7 mm) compared with the state-of-the-art (SOTA). Code is available at https://github.com/TangWen920812/TLT.

Keywords: Transformer · Cross attention · Registration

1 Introduction

The ability to accurately locate the location of follow-up lesions and subsequent quantitative assessment, referred to as "lesion tracking," is crucial to a variety of medical applications, in particular, cancer management. In practice, physicians need to spend significant time and effort to precisely match the same lesion across different time points. Thus, its investigation into a fully automated method of lesion tracking or lesion matching is highly desirable. However, compared with a large number of studies on lesion segmentation and detection [19,22], there are very few studies on lesion tracking [4,11]. In the field of natural images, there is a similar problem called target tracking or object tracking, for which several deep

learning-based methods have been proposed [3,12,23]. One of the simplest and most straightforward ideas is to apply these existing methods to lesion tracking tasks. However, lesion tracking is different from the aforementioned visual tracking in a number of aspects: (1) Medical imaging data are mostly in 3D format. (2) The lesion size varies at different time points, such as increasing, shrinking, or stabilizing. (3) The appearance of the lesion may change during the follow-up examination while its anatomical location remains unchanged. Thus, an effective lesion tracker should account for the differences in the lesion itself and be able to use anatomical information effectively. However, existing registration-based trackers [18,21] are not robust for small-sized lesions or heavily deformed lesions due to lack of sensitivity to local details, and Siamese networks [9,12] overlook the information around the lesion. Cai et al. [4] first provided an open-source dataset for lesion tracking and designed Deep Lesion Tracker (DLT), which combines the advantages of both strategies and obtained a baseline on this dataset. Although a large kernel size is extracted in cross correlation layers of DLT to encode the global image context, it is still susceptible to the inductive bias in convolution, leading to deviation in the aggregation of information around the lesion.

In this work, we leverage Transformer architecture, inspired by TransT [7], to replace existing cross correlation, and propose a novel Transformer Lesion Tracking framework (named TLT) using 3D features. To achieve our model, we design a Cross Attention-based Transformer (CAT) to capture global information. To better focus on useful features, we also introduce anatomical priors via the proposed Registration-based Anatomical Attention Module (RAAM) into CAT. Meanwhile, considering the memory cost in training process, we present a Sparse Selection Strategy (SSS) to extract the local effective information from the whole template image as input for CAT. In addition, we use a global regression as output to reduce the effect of insufficient multi-scale information and accelerate convergence. The experimental results show that the proposed method achieves better performance on the open-source dataset compared with the state-of-the-art methods.

2 Related Work

Registration-Based Trackers. The anatomy presented in a patient's medical images at different time points should be similar in the absence of surgery or similar treatment. Thus, lesion tracker should follow a spatial consistency that the tissue or the structure around a lesion, and the organ in which the lesion is located will not change significantly. Under this assumption, existing registration methods, such as Voxelmorph [1], provide solutions for lesion tracking. Since registration algorithms [10,15] focus on alignment or optimization on global structures, registration-based lesion trackers achieve decent performance for large-sized lesions or relatively stable lesions [18,21]. Still, due to the lack of sensitivity of registration algorithms to image details, these registration-based methods obtain reduced performance when dealing with small-sized lesions or heavily deformed lesions. In this study, we treat image registration as an auxiliary

operation, thereby improving model training efficiency as well as performance. Specifically, we select the mask registered via an affine registration method [15] as the prior attention and introduce it into the Transformer. The subsequent ablation experiment results demonstrate the effectiveness of this operation.

Siamese Networks. In recent years, Siamese-based methods have been popular in the field of visual object tracking. SiamFC [2], and its variants such as SiamRPN [13] and SiamRPN++ [12], are among the representative works. Subsequently, existing studies have demonstrated that lesion tracking could also be done using Siamese-based methods. Gomariz et al. [9] and Liu et al. [14] applied 2D Siamese networks for lesion tracking in ultrasound sequences. Whereas Rafael-Palou et al. [17] performed 3D Siamese networks to track lung nodules on CT series. Cai et al. [4] followed SiamRPN++ to use 3D Siamese networks to conduct universal lesion tracking in whole body CT images. These Siamese-based methods mainly consist of two parts: a backbone network for feature extraction and a correlation module to calculate the similarity between the template patch and the searching sub-region. However, such module is susceptible to the inductive bias of convolution operation and fails to fully utilize the global context, leading to local optimum in the optimization process. Thus, we introduce an attention-based Transformer architecture to focus on the key object in the global feature space, while replacing the correlation part.

Transformer-Based Tracking. In recent years, Transformer architecture [24] has taken over recurrent neural networks in natural language processing [8,20], and has also had an impact on the status of convolutional neural networks in computer vision [6,16]. More recently, Chen et al. [7] proposed a target tracking method on natural images with Transformer architecture instead of the cross correlation layers and achieved SOTA results. However, several issues remain to be addressed when applying Transformer to lesion tracking on 3D medical images. Specifically, to reduce memory cost and acquire features of different sized lesion adaptively, we design a sparse selection strategy to extract irregular patches from template feature maps as input to Transformer. To introduce prior anatomical structure information to Transformer, we create a registration-based attention guidance for auxiliary model training.

3   Method

Problem Description. Same as object tracking [2], lesion tracking aims to find its corresponding position in the searching image I_s when given a lesion in the template image I_t. Similar to [4], we simplify this task: given a lesion l in I_t with its known center c_t and radius r_t, we seek a mapping function $\mathcal{F}$ to predict the center c_s of l in I_s.

Overview. In this lesion tracking task, we define the baseline CT scan as the template image $I_t \in \mathbb{R}^{D_{t0} \times H_{t0} \times W_{t0}}$, and a corresponding follow-up CT scan as the searching image $I_s \in \mathbb{R}^{D_{s0} \times H_{s0} \times W_{s0}}$. D_{t0}, H_{t0} and W_{t0} represent the depth, height and width of the template image, respectively. And D_{s0}, H_{s0} and W_{s0} are

similarly defined for searching image. The proposed lesion tracking network (TLT) mainly consists of three components, as shown in Fig. 1. The feature extractor stacks 3D convolution and downsampling layers to efficiently represent the input volumes. The proposed sparse selection strategy is used for memory reducing and efficient feature acquisition. Then, the cross attention-based Transformer (CAT) is used to fuse the features of the searching and the selected template. In the CAT, a mask gained from the registration-based anatomical attention module (RAAM) is inserted to enhance fused features. Finally, the center predictor is responsible for getting the result from the output of Transformer.

3.1 Feature Extractor and Sparse Selection Strategy

In the proposed network, a modified 3D ResNet18 with shared weights is employed as the feature extractor. Compared with the original one, we remove the last stage of ResNet18, and take outputs of the fourth stage as final outputs. We also adjust the stride of the first convolutional layer from $2 \times 2 \times 2$ to $1 \times 1 \times 1$ to obtain a larger feature resolution. Considering the parameter redundancy and overfitting in 3D networks, the number of feature channels in each stage is reduced by half or more (see Fig. 1). Putting I_t and I_s through the learnable feature extractor respectively, their own image features $F_{t,ori} \in \mathbb{R}^{C \times D_t \times H_t \times W_t}$, $F_{s,ori} \in \mathbb{R}^{C \times D_s \times H_s \times W_s}$ are obtained for subsequent processes, where $D_t, H_t, W_t = \frac{D_{t0}}{8}, \frac{H_{t0}}{8}, \frac{W_{t0}}{8}$, $D_s, H_s, W_s = \frac{D_{s0}}{8}, \frac{H_{s0}}{8}, \frac{W_{s0}}{8}$, $C = 192$.

As shown in Fig. 1(a), template-based feature mining via the proposed **s**parse **s**election **s**trategy (SSS) precede the CTA to select features and to reduce memory cost. This is feasible because the location of the lesion in the template input I_t is known, and we believe the features $F_{t,ori}$ to have already incorporated local contextual information. The following are the details of the SSS flow. Given a lesion in the template image, we first generate a three-dimensional Gaussian map G based on the known center and radius of the lesion, which is formulated by:

$$G(c, r) = exp(-\frac{\sum_{i \in (x,y,z)} (i - c^i)^2}{\sum_{i \in (x,y,z)} (2r^i)^2})$$ (1)

Specifically, for the lesion l in I_t with its center c_t and radius r_t, the generated Gaussian map G_t is $G_t(c_t, r_t)$. Next, we resize G_t to the size of $F_{t,ori}$ by trilinear interpolation, and obtain the resized Gaussian map $\widetilde{G}_t$. Selecting a threshold Tr and using $\widetilde{G}_t$ as a reference mask, we extract valid features $F_{t,sparse}$ from the $F_{t,ori}$ as an input of the Transformer: $F_{t,sparse} = F_{t,ori}(x, y, z | \widetilde{G}_t(x, y, z) > Tr)$

3.2 Cross Attention-Based Transformer

Unlike the similarity-based correlation module used in the previous Siamese-based networks, we design a Cross Attention-based Transformer (CAT) to combine global and local context. Queries Q, keys K and values V are encoded from same source in Transformer [24]. But in CAT, to grab global context and blend multiple features of different sizes, we adopt cross-attention (CA), in which K, V
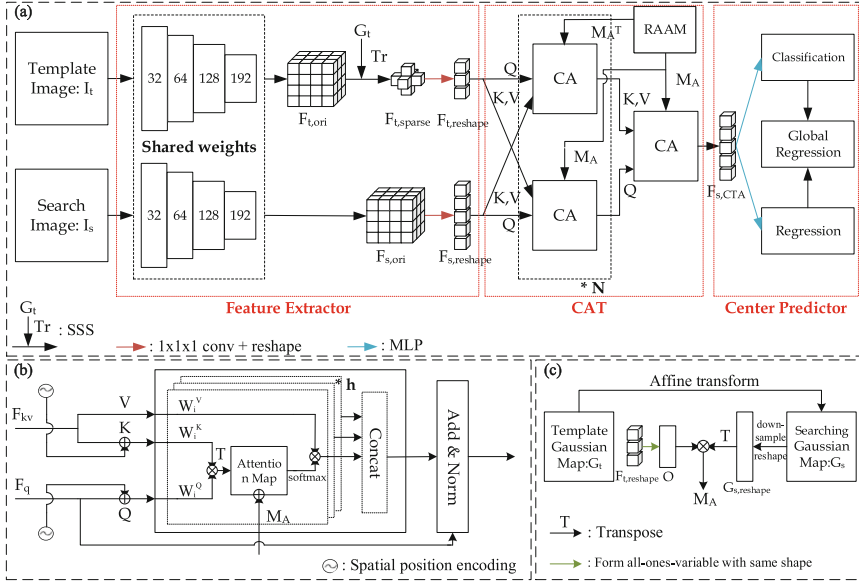
Fig. 1. (a) Overall structure of the proposed network. (b) Cross attention in CAT. (c) Structure of RAAM

are stemmed from the same input while Q from another. In TLT, we put a CA on each of the template and searching path respectively (see Fig 1(a)). In CA of the template line, K and V are encoded from the reshaped $F_{s,ori} : F_{s,reshape} \in \mathbb{R}^{D_s H_s W_s \times C}$ and Q is encoded from the reshaped $F_{t,sparse} : F_{t,reshape} \in \mathbb{R}^{L \times C}$, $L = \text{len}(F_{t,sparse})$. While in CA of the searching line, K, V and Q have the opposite origins to those in the template one. In short, as shown in Fig. 1(b), Q is encoded from the features which need enhancement (F_q), and K, V are encoded from the other (F_{kv}). We apply CA on each lines for $N(N = 3)$ times, and use another CA on searching line to obtain the final output features: $F_{s,\text{CTA}}$.

We further create a novel structure called Registration-based Anatomical Attention Module (RAAM) to calculate an anatomical information mask M_A, whose transpose is taken as M_A^{T} (see Fig. 1(c)). As described above, the anatomical information is needed for lesion tracking. Thus, we create a matrix to provide the anatomical information for each of template and searching side. For template side, we assume all the voxels in $F_{t,reshape}$ are of the same importance, and we build a matrix $O \in \mathbb{R}^{L \times 1}$, in which all elements are 1. For searching side, we first use an affine registration method [15] to roughly align I_t and I_s by solving: $\mathcal{T}_{\text{Aff}} = \arg \min \|\mathcal{T}_{\text{Aff}}(I_t) - I_s\|_1$. We choose to use an affine registration method instead a non-rigid one because the non-rigid registration is slow and provides restriction to the attention that limit the model's ability to learn for local variation and details. Then, we can obtain a registration-based Gaussian map $G_s = \mathcal{T}_{\text{Aff}}(G_t)$. Afterwards, we downsample G_s to the size of $F_{s,ori}$ and reshape

it to $G_{s,reshape} \in \mathbb{R}^{D_s H_s W_s \times 1}$, which is defined as the matrix of searching side. Finally, M_A can be calculated by the following formula:

$$M_A = O \otimes G_{s,reshape}^{\mathrm{T}} \tag{2}$$

where $\otimes$ is matrix multiplication operation, and $M_A \in \mathbb{R}^{L \times D_s H_s W_s}$. So the attention we use in CAT at each head (see Fig. 1(b)) can be define as following:

$$\text{Attention}_i(Q, K, V) = \text{softmax}\left(\frac{(QW_i^Q)(KW_i^K)^{\mathrm{T}}}{\sqrt{d_k}} + M_A\right)(VW_i^V), \tag{3}$$

where W_i^Q, W_i^K, W_i^V are parameter matrices, d_k is the dimension of key, $i \in \{1, ..., h\}$ is the index of head and h is the number of heads in multiple head attention.

3.3   Center Predictor and Training Loss

Similar to the head of detection networks, our center predictor consists of a classification branch and a regression branch, where each branch is a multilayer perceptron (MLP). The classification head is to classify if a voxel from the output is inside of a lesion, and the regression head is to regress the exact center position. In detail, after inputting the features $F_{s,\text{CTA}}$, the predictor outputs the classification results $\hat{Y} \in \mathbb{R}^{1 \times D_s H_s W_s}$ and center coordinates $\hat{C} \in \mathbb{R}^{3 \times D_s H_s W_s}$. During training, we define the ground truth as a Gaussian map generated by Eq. 1 with the target center c_s and the corresponding radius r_s. We downsample it to obtain the Gaussian label G_L which matches the size of $\hat{Y}$, and obtain label $Y = \frac{G_L - min(G_L)}{max(G_L) - min(G_L)}$. L1 loss is used as the regression loss, which is formulated as:

$$L_r = ||\hat{c} - c_s||_1, \ \hat{c} = \sum \text{softmax}(\hat{Y}) * \hat{C} \tag{4}$$

where $\hat{c}$ is the final output of the center predictor, which we define as global regression. Meanwhile, a focal loss [4] is used as the classification loss for auxiliary training:

$$L_c = \sum_i \begin{cases} (1 - \hat{y}_i)^\alpha \log(\hat{y}_i) & \text{if } y_i = 1 \\ (1 - y_i)^\beta (\hat{y}_i)^\alpha \log(1 - \hat{y}_i) & \text{otherwise} \end{cases} \tag{5}$$

where y_i and $\hat{y}_i$ are the i-th elements in Y and $\hat{Y}$, respectively, and $\alpha = \beta = 2$.

4   Experiments and Experimental Results

4.1   Dataset and Experiment Setup

Dataset. We validate our method on a public dataset, DLS [4], which consists of CT image pairs inherited from DeepLesion [27]. There are 3008, 403 and 480 lesion pairs for training, validation, and testing in this dataset, respectively.

Table 1. Lesion tracking comparison on Deep Lesion Tracking testing dataset. * represents the p value of paired t-test is smaller than 0.05.

Method	CPM@ 10 mm	CPM@ Radius	MED$_X$ (mm)	MED$_Y$ (mm)	MED$_Z$ (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 [1]	49.90	65.59	4.6 ± 6.7	5.2 ± 7.9	6.6 ± 6.2	10.9 ± 10.9
LENS-LesioGraph [25, 28]	63.85	80.42	$\mathbf{2.6 \pm 4.6}$	2.7 ± 4.5	6.0 ± 8.6	8.0 ± 10.1
VULD-LesionGraph [5, 28]	64.69	76.56	3.5 ± 5.2	4.1 ± 5.8	6.1 ± 8.8	9.3 ± 10.9
VULD-LesaNet [5, 26]	65.00	77.81	3.5 ± 5.3	4.0 ± 5.7	6.0 ± 8.7	9.1 ± 10.8
SiamRPN++ [12]	68.85	80.31	3.8 ± 4.8	3.8 ± 4.8	4.8 ± 7.5	8.3 ± 9.2
LENS-LesaNet [25, 26]	70.00	84.58	2.7 ± 4.8	$\mathbf{2.6 \pm 4.7}$	5.7 ± 8.6	7.8 ± 10.3
DEEDS [10]	71.88	85.52	2.8 ± 3.7	3.1 ± 4.1	5.0 ± 6.8	7.4 ± 8.1
DLT-Mix [4]	78.65	88.75	3.1 ± 4.4	3.1 ± 4.5	4.2 ± 7.6	7.1 ± 9.2
DLT [4]	78.85	86.88	3.5 ± 5.6	2.9 ± 4.9	4.0 ± 6.1	7.0 ± 8.9
TransT [7]	79.59	88.99	3.4 ± 5.9	5.4 ± 6.1	1.8 ± 2.2	7.6 ± 7.9
TLT	$\mathbf{87.37^*}$	$\mathbf{95.32^*}$	3.0 ± 6.2	3.7 ± 5.2	$\mathbf{1.7 \pm 2.1}$	$\mathbf{6.0 \pm 7.7^*}$

Since the ground truth lesion center of all lesions in this dataset and the corresponding radius are defined, we could mutually track within a lesion pair. Therefore, a total of 906 and 960 directed lesion pairs are used for evaluation in validation and testing sets, respectively.

Evaluation Metrics. The center point matching (CPM) accuracy is selected to evaluate the performance of lesion matching. As defined in [4], a match will be counted correct when the Euclidean distance between ground truth and predicted centers is smaller than a threshold (@10 mm: min(10 mm, r_s), @Radius: r_s). The mean Euclidean distance (MED) in mm $+/-$ standard deviation between ground truth and predicted centers, and its projections in each direction (denoted as MED$_X$, MED$_Y$ and MED$_Z$, respectively) [4] are also used for model evaluation.

Implementation Details. The proposed method is implemented using PyTorch (v1.5.1). The network is optimized by Adam with initial learning rate of 0.0001 and trained for 300 epochs. The batch size is 4 and the number of parameters of the model is 5.98M. All CT volumes have been resampled to the isotropic resolution of 1 mm before feeding into the network. This training setting is used in all deep learning-based methods selected for comparison. For the affine registration method [15] and DEEDS [10], following the setting of [4] and [10], all CT volumes are resampled to an isotropic resolution of 2 mm.

4.2   Experimental Results and Discussion

Model Comparison. We took TransT [7] as baseline, and selected DLT and other state-of-the-art approaches in [4] for comparison. Table 1 shows the quantitative results of these methods. Our method yields a CPM@10 mm of 87.37, a CPM@*Radius* of 95.32, and a MED of 6.0 ± 7.7, which outperforms all the

Table 2. Ablation study on each module and different thresholds. * represents the p value of paired t-test is smaller than 0.05.

SSS	RAAM	Global Regressor	CPM@ 10 mm	MED (mm)	Threshold	CPM@10 mm	MED(mm)
					0.9	83.57	6.80 ± 8.12
					0.8	83.99	6.65 ± 7.99
			79.59	7.58 ± 7.91	0.7	87.37^*	$\mathbf{5.98 \pm 7.68}^*$
✓			84.78	6.76 ± 7.86	0.6	86.70	6.26 ± 7.88
✓	✓		86.58	6.30 ± 7.79	0.5	86.37	6.20 ± 7.83
✓	✓	✓	$\mathbf{87.37}^*$	$\mathbf{5.98 \pm 7.68}^*$	0.4	85.01	6.39 ± 7.98

compared methods in terms of both CPM and MED metrics. A paired t-test is used on CPM@10 mm, CPM@$Radius$ and MED to perform statistical tests. Moreover, we observe that transformer-based methods, TransT and our TLT, both achieving large improvements in terms of MED_Z compared with methods that use convolution to compute similarity. This may be because the Transformer focuses more on the information in the z-axis direction, which is also consistent with physician cognition.

Ablation Study. To evaluate the effectiveness of various configurations in our proposed method, we conduct ablation experiments from two aspects: module setting and threshold setting. A paired t-test is also used for statistical tests. Table 2 shows the experimental results. The results show that accuracy drops with each module change, which validates the competence of our proposed method. Meanwhile, it is observed that the threshold of 0.7 is much better than that of other thresholds. Therefore, we choose 0.7 as the thresholds Tr in our TLT.

Discussion. As we observe, in ablation study, the SSS module leads to the biggest improvement. To verify this, we also perform ablation study with only one single module removed, as shown in Table 1 in supplementary materials. This happens when there are many small lesions in the dataset, such as lung nodules. If these small lesions are cropped on the original image, due to downsampling, the feature map will become very small, and in the last several downsampling processes will always become one voxel, which could lead to a decline in performance. The SSS solves this problem by selecting voxels on the last feature map. Even if only one voxel on the feature map is selected, this voxel can still obtain more surrounding information in the networks than without SSS. Meanwhile, based on our observations, we found that when the registration method failed, sometimes our model would fail as well. This is because we use registration to feed anatomical information to the transformer, and anatomical information helps the transformer accelerate convergence, which forms a dependency. In addition, when there are similar lesions in similar locations, such as two solid nodules at the edge of the right upper lung, and only a few layers difference in the z-axis direction, the model will also be confused.

5   Conclusion

This paper presents a novel Transformer-based framework for lesion tracking by leveraging both the anatomical prior and the cross image relevance. We further introduce a global regression to integrate multi-scale information while using sparse selection strategy to reduce memory consumption. TLT achieves the state-of-the-art performance on DLT dataset, significantly exceeding previous methods in lesion tacking accuracy. Future work includes multi-institutional validation and reader studies to examine the efficiency improvement for physicians in clinical setting.

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9252–9260 (2018)
2. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: Fully-convolutional Siamese networks for object tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 850–865. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_56
3. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: 2010 IEEE Computer Society Conference on Computer Vision And Pattern Recognition, pp. 2544–2550. IEEE (2010)
4. Cai, J., et al.: Deep lesion tracker: monitoring lesions in 4d longitudinal imaging studies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15159–15169 (2021)
5. Cai, J., et al.: Deep volumetric universal lesion detection using light-weight pseudo 3D convolution and surface point regression. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 3–13. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_1
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
7. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8126–8135 (2021)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Gomariz, A., Li, W., Ozkan, E., Tanner, C., Goksel, O.: Siamese networks with location prior for landmark tracking in liver ultrasound sequences. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1757–1760. IEEE (2019)

10. 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**(7), 1239–1248 (2013)

11. Hering, A., et al.: Whole-body soft-tissue lesion tracking and segmentation in longitudinal CT imaging studies. In: Medical Imaging with Deep Learning, pp. 312–326. PMLR (2021)

12. Li, B., et al.: Evolution of Siamese visual tracking with very deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 16–20 (2019)

13. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with Siamese region proposal network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8971–8980 (2018)

14. Liu, F., Liu, D., Tian, J., Xie, X., Yang, X., Wang, K.: Cascaded one-shot deformable convolutional neural networks: developing a deep learning model for respiratory motion estimation in ultrasound sequences. Med. Image Anal. **65**, 101793 (2020)

15. Marstal, K., Berendsen, F., Staring, M., Klein, S.: Simpleelastix: a user-friendly, multi-lingual library for medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 134–142 (2016)

16. Parmar, N., et al.: Image transformer. In: International Conference on Machine Learning, pp. 4055–4064. PMLR (2018)

17. Rafael-Palou, X., et al.: Re-identification and growth detection of pulmonary nodules without image registration using 3d Siamese neural networks. Med. Image Anal. **67**, 101823 (2021)

18. Raju, A., et al.: Co-heterogeneous and adaptive segmentation from multi-source and multi-phase CT imaging data: a study on pathological liver and lesion segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12368, pp. 448–465. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58592-1_27

19. Shao, Q., Gong, L., Ma, K., Liu, H., Zheng, Y.: Attentive CT lesion detection using deep pyramid inference with multi-scale booster. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 301–309. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_34

20. Synnaeve, G., et al.: End-to-end ASR: from supervised to semi-supervised learning with modern architectures. arXiv preprint arXiv:1911.08460 (2019)

21. Tan, M., et al.: A new approach to evaluate drug treatment response of ovarian cancer patients based on deformable image registration. IEEE Trans. Med. Imaging **35**(1), 316–325 (2015)

22. Tang, W., et al.: M-SEAM-NAM: multi-instance self-supervised equivalent attention mechanism with neighborhood affinity module for double weakly supervised segmentation of COVID-19. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12907, pp. 262–272. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87234-2_25

23. Teed, Z., Deng, J.: RAFT: recurrent all-pairs field transforms for optical flow. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 402–419. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_24

24. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 1–7 (2017)

25. Yan, K., et al.: Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in CT. IEEE Trans. Med. Imaging **40**(10), 2759–2770 (2020)
26. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8523–8532 (2019)
27. Yan, K., Wang, X., Lu, L., Summers, R.M.: DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J. Med. Imaging **5**(3), 036501 (2018)
28. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9261–9270 (2018)