

Accurate and Robust Lesion RECIST Diameter Prediction and Segmentation with Transformers

Youbao Tang^{1✉}, Ning Zhang¹, Yirui Wang¹, Shenghua He¹, Mei Han¹, Jing Xiao², and Ruei-Sung Lin¹

¹ PAII Inc., Palo Alto, CA, USA
tybxiaobao@gmail.com

² Ping An Technology, Shenzhen, China

Abstract. Automatically measuring lesion/tumor size with RECIST (Response Evaluation Criteria In Solid Tumors) diameters and segmentation is important for computer-aided diagnosis. Although it has been studied in recent years, there is still space to improve its accuracy and robustness, such as (1) enhancing features by incorporating rich contextual information while keeping a high spatial resolution and (2) involving new tasks and losses for joint optimization. To reach this goal, this paper proposes a transformer-based network (MeaFormer, **M**ea**Form**er **transForm**er) for lesion RECIST diameter prediction and segmentation (LRDPS). It is formulated as three correlative and complementary tasks: lesion segmentation, heatmap prediction, and keypoint regression. To the best of our knowledge, it is the first time to use keypoint regression for RECIST diameter prediction. MeaFormer can enhance high-resolution features by employing transformers to capture their long-range dependencies. Two consistency losses are introduced to explicitly build relationships among these tasks for better optimization. Experiments show that MeaFormer achieves the state-of-the-art performance of LRDPS on the large-scale DeepLesion dataset and produces promising results of two downstream clinic-relevant tasks, *i.e.*, 3D lesion segmentation and RECIST assessment in longitudinal studies.

Keywords: RECIST diameter prediction · Lesion segmentation · Transformers · Keypoint regression.

1 Introduction

When reading computed tomography (CT) scans, an important step for radiologists is to measure the size of found lesions. Accurately measuring the lesions is crucial in precisely assessing the lesion growth rates across different time points. As such, it plays an important role in monitoring disease progression and making therapeutic plan. Currently, the most widely-used clinical guideline of lesion measurement is bidimensional RECIST (Response Evaluation Criteria In Solid Tumors) [10] diameters. Manually labeling the RECIST diameters is tedious, time-consuming, and prone to be inconsistent among different observers [25].

To overcome these issues, Tang *et al.* [19] first presented a semi-automatic RECIST diameter labeling approach. It has been demonstrated that segmentation can be considered as a more precise measurement than RECIST diameters to assess treatment response [16,3]. Therefore, many approaches [4,17,2,1,21,18,20] have been proposed for lesion segmentation. Meanwhile, some great progresses [21,20] have been made recently to further improve the performance of RECIST diameter prediction and incorporate lesion segmentation simultaneously.

The previous works [21,20] tried to build powerful convolutional neural networks (CNN) through sophisticated designs to learn discriminative features. Although they achieved promising performances on lesion RECIST diameter prediction and segmentation (LRDPS), there are two obvious limitations. First, they failed to fully exploit the features' long-range dependencies due to the inherent locality of convolutional operations, which is crucial for pixel-wise prediction tasks, *e.g.*, lesion segmentation and heatmap prediction. Second, they only produced a single RECIST diameter output based on heatmap prediction. This would lead to unsatisfactory results when the lesions are difficult to identify.

On the other hand, transformer is designed to model sequence-to-sequence predictions in natural language processing tasks [22], which is able to capture the long-range dependencies of the input sequences. Recently, it has been successfully applied to tasks of computer vision [9,7] and medical image segmentation [6,13,26], which attracts tremendous attentions increasingly. Hence, to address the above limitations, we propose a transformer-based network (MeaFormer, **M**easurement **transFORMer**) for LRDPS in this paper, where three correlative and complementary tasks are involved in including lesion segmentation, heatmap prediction, and keypoint regression. Three main contributions are made in MeaFormer to push it into achieving accurate and robust results of LRDPS: (1) transformers are employed to model the long-range dependencies of the high-resolution features extracted by a small CNN backbone. By incorporating rich contextual information, the enhanced high-resolution features facilitate accurate lesion segmentation and heatmap prediction results. (2) To the best of knowledge, it is the first time to use keypoint regression for RECIST diameter prediction and obtain RECIST diameters directly from segmentations. Thus, three types of RECIST diameter predictions are produced by MeaFormer in total instead of a single one. Fusing them helps to achieve a more robust outcome. (3) Two consistency losses are introduced to explicitly build relationships among these tasks. A joint optimization using them facilitates MeaFormer with an more optimal solution. Furthermore, extensive experiments on the large-scale DeepLesion dataset show that the proposed MeaFormer achieves the state-of-the-art performance of LRDPS.

2 Method

As done in [21,20], we also develop a click-based automatic LRDPS system with two steps, whose pipeline is presented in Fig. 1(a). Given a CT image and a click guidance (the red spot) within the lesion, the top-left and bottom-right corners

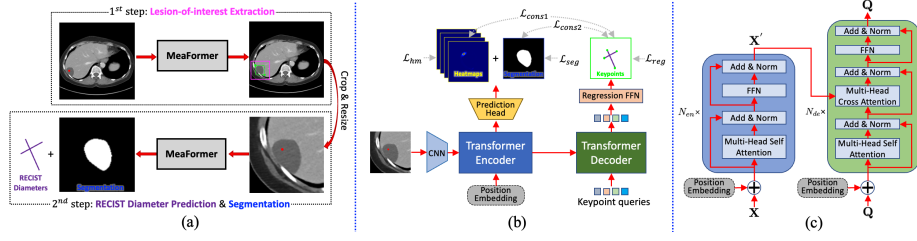


Fig. 1. System overview. (a) The pipeline of our click-based automatic LRDPS system with separately trained MeaFormer models for step 1 and step 2, respectively. (b) The architecture of the proposed network MeaFormer used at step 2. (c) The detailed structures of MeaFormer’s encoder and decoder. The MeaFormer used at step 1 has the same architecture except the following differences: (i) two queries for box prediction at step 1 rather than four at step 2 and (ii) no consistency losses (*i.e.*, \mathcal{L}_{cons1} and \mathcal{L}_{cons2}) used for training at step 1. All figures are best viewed in color.

(the green circles) of a bounding box (the green dotted box) are predicted to indicate the lesion region at the 1st step. A lesion-of-interest (LOI, the region in the magenta box) is extracted, whose center is the predicted box and whose width is two times the extent of the predicted box’s long side. As such, sufficient context information is preserved in a LOI. The LOI is then cropped and resized to 256×256 . At the 2nd step, the results of LRDPS are obtained based on the resized LOI. The proposed network MeaFormer is used at both steps.

2.1 The Architecture of MeaFormer

Fig. 1(b) shows the architecture of MeaFormer. We borrowed idea from DETR [7] where transformers are applied on a CNN-based feature map instead of the image input. MeaFormer contains four main components, including a CNN backbone for feature extraction, an encoder-decoder transformer, a CNN-based prediction head for lesion segmentation and heatmap prediction, and a regression feed-forward network (FFN) for box/keypoint prediction at the 1st/2nd step.

Backbone. Given an image $\mathbf{I}_{ct} \in \mathbb{R}^{H_0 \times W_0}$ of an original CT image at the 1st step or a resized LOI CT image at the 2nd step, we first generate a click image \mathbf{I}_c and a distance transform image \mathbf{I}_d based on the human click guidance following [21, 20]. A 3-channel image $\mathbf{I} \in \mathbb{R}^{3 \times H_0 \times W_0}$ is constructed by concatenating \mathbf{I}_{ct} , \mathbf{I}_c , and \mathbf{I}_d , which serves as an input of the backbone. To reduce the computation cost at the 1st step, the original CT image is resized to 256×256 . Thus, $H_0 = W_0 = 256$ in this work. Here, we choose HRNet-W48 [23] as the backbone for feature extraction. As we know that the features’ spatial and low-level information are crucial to getting accurate pixel-wise predictions, thus only the first three stages of HRNet-W48 is adopted to generate a low-level and high-resolution feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, where $C = 48$, $H = \frac{H_0}{4}$, $W = \frac{W_0}{4}$.

Encoder-Decoder Transformer. We use a vanilla transformer architecture [22] for our transformer encoder and decoder design, whose detailed structures

are shown in Fig. 1(c). Since the encoder takes as input a sequence, the feature map \mathbf{F} is flattened and permuted to a sequence $\mathbf{X} \in \mathbb{R}^{(H \cdot W) \times C}$. In this work, the encoder has $N_{en} = 6$ layers. Each encoder layer contains a multi-head self-attention module and a feed-forward network (FFN). To make use of the spatial order of the sequence, the fixed positional encodings are added to the input sequence of each encoder layer. A new sequence $\mathbf{X}' \in \mathbb{R}^{(H \cdot W) \times C}$ is produced after forwarding N_{en} encoder layers. The decoder aims to transform a learnable query matrix $\mathbf{Q} \in \mathbb{R}^{N_q \times C}$ that contains N_q query embeddings of size C using multi-head self- and encoder-decoder attention mechanisms. It consists of $N_{de} = 6$ layers. Each layer takes as input \mathbf{X}' and \mathbf{Q} , and outputs an updated \mathbf{Q} . Similarly to the encoder, the positional encodings are added to the input of each decoder layer. After forwarding N_{de} decoder layers, the final updated query matrix \mathbf{Q} is taken as input of the regression FFN. We set $N_q = 2$ and $N_q = 4$ at the 1st and 2nd step, respectively. That is because two corners/four keypoints are required for box/keypoint prediction.

Prediction Head. A CNN-based prediction head is attached to the transformer encoder for lesion segmentation and heatmap prediction. Before feeding the sequence $\mathbf{X}' \in \mathbb{R}^{(H \cdot W) \times C}$ into the prediction head, we first permute and reshape \mathbf{X}' to a feature map $\mathbf{F}' \in \mathbb{R}^{C \times H \times W}$. The prediction head consists of sequential layers: $conv(kn = 32, ks = 3, st = 1)$, $deconv(32, 4, 2)$, $conv(32, 3, 1)$, $deconv(32, 4, 2)$, and $conv(5, 1, 1)$, where $conv(*)$, $deconv(*)$, kn , ks , and st represent convolutional layer, deconvolutional layer, kernel number, kernel size, and stride, respectively. Each layer is followed by a batch normalization and a ReLU except the last one. Therefore, the output of the prediction head $\mathbf{O} \in \mathbb{R}^{5 \times H_0 \times W_0}$ has the same resolution to the input image. One channel of \mathbf{O} is the lesion segmentation result \mathbf{S} and the other four channels are the predicted heatmaps \mathbf{M} .

Regression FFN. To get the box/keypoint prediction results, a regression FFN is attached to the transformer decoder. It consists of a 2-layer perceptron with ReLU activation function and hidden dimension $d = 96$, and a linear projection layer with two output nodes for predicting the corner/keypoint coordinate (x, y) using a softmax function. As done in [7], we also add the regression FFN after each transformer decoder layer to introduce auxiliary losses for model optimization. All regression FFNs share their parameters.

2.2 Model Optimization

The proposed MeaFormer has three task-related losses to optimize. The lesion segmentation loss \mathcal{L}_{seg} is the summation of a binary cross entropy loss \mathcal{L}_{bce} and an IoU loss \mathcal{L}_{iou} [15], *i.e.*, $\mathcal{L}_{seg} = \mathcal{L}_{bce} + \mathcal{L}_{iou}$. The same procedure as [18] is conducted to construct initial and updated lesion pseudo masks as supervision for \mathcal{L}_{seg} optimization. For heatmap prediction, the objective function \mathcal{L}_{hm} is a mean squared error loss \mathcal{L}_{mse} that measures the errors between the predicted heatmaps and the ground truth heatmaps, which are four 2D Gaussian maps (with a standard deviation of 5 pixels) centered on the endpoints of RECIST annotations. For keypoint regression, the objective function \mathcal{L}_{reg} is a L1 loss \mathcal{L}_1

that is adopted to calculate the errors between the predicted keypoint coordinates produced by N_{de} regression FFNs and the ground truth coordinates of the RECIST diameters' endpoints.

Besides these individual losses, this work also introduces two consistency losses to explicitly build two relationships among different tasks. One relationship is built between the tasks of keypoint regression and heatmap prediction. Both of them are conducted for RECIST diameter prediction, thus their outputs should be as consistent as possible. Let (x_i, y_i) denote the i^{th} keypoint's predicted coordinate produced by the keypoint regression task, so \mathbf{M}_{i,y_i,x_i} means the value of the i^{th} predicted heatmap at location (x_i, y_i) . To make the outputs of keypoint regression and heatmap prediction as consistent as possible, (x_i, y_i) should be as close as possible to the maximum location of the i^{th} predicted heatmap, meaning that the error between \mathbf{M}_{i,y_i,x_i} and 1 should be as small as possible. The other relationship is built between the tasks of keypoint regression and lesion segmentation. In an ideal situation, the predicted keypoints should be located at the boundaries of the lesion segmentation. We first binarize the segmentation result by $\mathbf{S} \geq 0.5$, and then perform a distance transformation on the binary mask to get a distance map \mathbf{D} . Therefore, to make the results of keypoint regression and lesion segmentation as consistent as possible, the error between \mathbf{D}_{y_i,x_i} and 0 should be as small as possible. L1 losses are used to compute the aforementioned two errors, which are denoted as \mathcal{L}_{cons1} and \mathcal{L}_{cons2} , respectively. The final objective function for model optimization is defined as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{seg} + \lambda_2 \mathcal{L}_{hm} + \lambda_3 \mathcal{L}_{reg} + \lambda_4 \mathcal{L}_{cons1} + \lambda_5 \mathcal{L}_{cons2}, \quad (1)$$

where we set $\lambda_1 = 1$, $\lambda_2 = 10$, $\lambda_3 = 1$, $\lambda_4 = 0.01$, and $\lambda_5 = 0.01$ to balance the magnitude of different losses in this work.

Implementation Details. MeaFormer is implemented in PyTorch 1.6 [14] and the CNN backbone is initialized with ImaeNet [8] pre-trained weights. Adam optimizer [12] with an initial learning rate of 0.001 and default settings (*e.g.*, $\text{betas}=(0.9, 0.999)$) is used to train the model for 200 epochs reduced by 0.1 at epoch 100 and 150. The training batch size is 16. The following operations are randomly conducted for data augmentation: scaling, cropping, rotating, brightness and contrast adjusting, and Gaussian blurring. The MeaFormer models used at two steps are trained separately. Following [18], the one at the 2^{nd} step is trained three rounds with the iteratively updated lesion pseudo masks as supervisions for weakly-supervised lesion segmentation learning.

3 Experiments

Dataset and Evaluation Metrics. A large-scale DeepLesion dataset [24]¹ is used to train and test the proposed MeaFormer. It collects 32,735 lesions from 10,594 studies of 4,427 patients. There are a variety of lesion types in this dataset, such as lung nodules, liver tumors, enlarged lymph nodes, and so on. All

¹ <https://nihcc.app.box.com/v/DeepLesion>

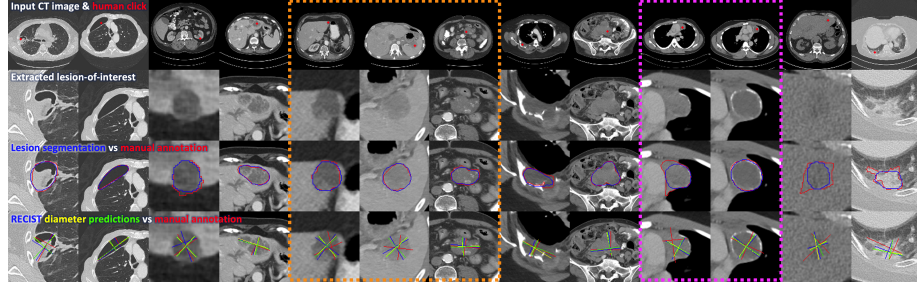


Fig. 2. Qualitative results of our system. Each column gives an example, where the content of each row has been described with texts. The predicted RECIST diameters shown in the fourth row with different colors are obtained from three different tasks, including **segmentation**, **heatmap prediction**, and **keypoint regression**. The orange dotted box shows three examples with imperfect manual diameter annotations. The magenta dotted box shows a lesion pair collected from the same patient at different time points, where the manual annotations of the left lesion are incorrect.

Table 1. Results of different methods for LRDPS on the DeepLesion test set in terms of model size (*i.e.*, the parameter number, the unit is million), Dice (%), and long/short diameter length error (mm). The mean and standard deviation are reported. **Red** and **blue** texts indicate rank 1 and 2, respectively.

Method	Size	Dice	Heatmaps		Segmentation		Regression		Fusion	
			Long	Short	Long	Short	Long	Short	Long	Short
Cai <i>et al.</i> [4]	-	90.6±8.9	-	-	-	-	-	-	-	-
AHRNet [18]	-	92.6±4.3	-	-	-	-	-	-	-	-
Tang <i>et al.</i> [19]	-	-	1.9±2.2	1.6±1.9	-	-	-	-	-	-
SEENet [21]	-	91.2±3.9	1.7±2.0	1.6±1.8	-	-	-	-	-	-
nnUNet [11]	-	90.7±5.0	2.1±2.0	1.8±1.7	-	-	-	-	-	-
PDNet [20]	-	92.4±4.5	1.7±1.5	1.5±1.4	-	-	-	-	-	-
Swin-Unet [6]	84.0	91.0±5.1	1.9±1.9	2.0±3.7	2.0±1.8	1.8±2.2	-	-	1.9±1.8	1.7±2.5
SegTran [13]	134.8	91.9±5.2	1.8±1.5	1.9±3.1	1.8±1.7	1.7±2.1	-	-	1.7±1.6	1.7±2.2
TransFuse [26]	26.1	92.0±4.9	1.8±1.4	1.8±2.0	1.7±1.6	1.5±2.1	-	-	1.7±1.5	1.6±1.8
Ours	18.3	92.7±4.3	1.6±1.4	1.4±1.6	1.6±1.3	1.5±1.5	1.9±1.5	1.7±1.9	1.6±1.3	1.4±1.5

lesion are annotated by bidimensional RECIST diameters telling their sizes and locations. As done in [21,20], 1,000 lesion images from 500 patients are manually segmented as a test set. The rest patient data are used for training (80%) and validation (20%). The same evaluation metrics to [21,20] are adopted to calculate the quantitative results, including the pixel-wise dice coefficient (Dice) for lesion segmentation and the differences between the diameter lengths (mm) of the predictions and manual annotations for RECIST diameter prediction.

Experimental Results. Fig. 2 displays several qualitative results produced by our system. We observe that (1) the lesions are well located at the center of the extracted LOIs, meaning that the proposed network can accurately predict the corners of the lesion bounding boxes. A prediction is considered as correct when the IoU of the predicted box and the ground truth box is larger than 0.5. On the 1000 test lesions, we achieve an accuracy of 99.1% surpassing the previous work

Table 2. Results of different settings of our network on the DeepLesion test set.

Method	Size	Dice	Heatmaps		Regression	
			Long	Short	Long	Short
HRNet _{regression}	39.2	-	-	-	3.0±2.3	2.1±2.2
HRNet _{heatmap}	70.4	91.4±5.7	1.9±1.7	1.9±2.4	-	-
HRNet*+TranE	17.6	92.1±4.9	1.7±1.6	1.7±2.1	-	-
HRNet*+TranE+TranD(\mathcal{L}_1)	18.3	92.5±4.4	1.7±1.7	1.6±1.7	2.3±1.9	2.0±2.2
HRNet*+TranE+TranD($\mathcal{L}_1+\mathcal{L}_{cons1}$)	18.3	92.5±4.5	1.7±1.5	1.5±1.7	1.9±1.6	1.8±2.0
HRNet*+TranE+TranD($\mathcal{L}_1+\mathcal{L}_{cons1}+\mathcal{L}_{cons2}$)	18.3	92.7±4.3	1.6±1.4	1.4±1.6	1.9±1.5	1.7±1.9

[21] by about 2%. (2) Our automatic results are close to the manual annotations and the diameter predictions produced by different tasks are nearly consistent, which visually demonstrate the effectiveness of our system and the contributions of the proposed consistency losses. (3) For some cases (*e.g.*, the examples in the orange dotted box), even the manual diameter annotations are not very precise, our automatic predictions can mitigate this issue. (4) For some cases (the left example in the magenta dotted box), the lesions' boundaries are unclear for some areas, resulting in incorrect manual annotations, but the automatic results can correct them. (5) Although our system works well on most of cases, it still faces some difficulties when the lesions' boundaries/shapes are heavily vague/irregular (the examples in the last two columns).

The quantitative comparisons of our system with previous approaches on these tasks [4,19,11,21,18,20] and three transformer-based medical image segmentation approaches [6,13,26] are presented in Table 1. The listed results of [4,19,21,18,20] are copied from their related papers. The results of nnUnet [11] are copied from [20]. For [6,13,26], we train them for heatmap prediction and lesion segmentation at the 2nd step using the same data as ours for fair comparisons. We observe that (1) Our MeaFormer has the smallest model size but achieves the best performance on all tasks, demonstrating its effectiveness for these tasks. (2) Better segmentation results produce better RECIST diameter predictions from segmentation. Sometimes, it even outperforms the heatmap-based prediction. This demonstrates that segmentation is a reliable way to predict RECIST diameters. Specifically, the RECIST diameters are obtained by first finding the longest diameter (long axis) from a lesion segmentation, and then calculating its longest perpendicular diameter (short axis). (3) More reliable RECIST diameter lengths can be obtained after fusing the predictions from different tasks, which mimics the behavior of radiologists making a consistent decision together when they have different observations. In this work, we find that the following fusion strategy achieves the best performance: (i) use prediction from segmentation as the reference, (ii) select a prediction from another task that has the smallest length difference compared to the reference, (iii) average their lengths to estimate the lesion size.

Ablation Studies. The proposed network is built based on HRNet by gradually integrating a transformer encoder (TranE) and a transformer decoder (TranD). Besides, two new consistency losses (\mathcal{L}_{cons1} and \mathcal{L}_{cons2}) are introduced for model optimization. Although only the first three stages of HRNet (HRNet*) are used in

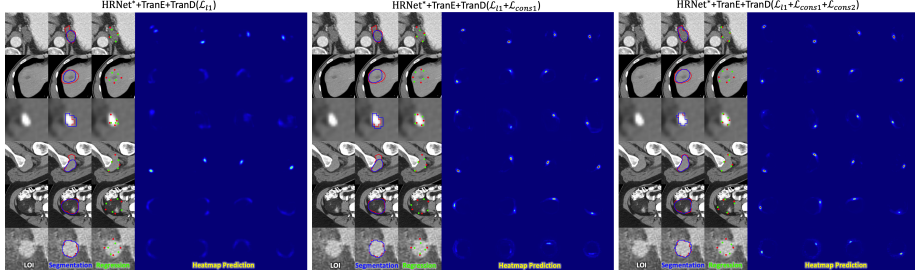


Fig. 3. Visual comparisons of the results produced by the proposed network that is trained with or without the consistency losses. The red curves and spots are the **manually annotated lesion masks and RECIST diameters' endpoints**. The blue curves are the **automatic lesion segmentation results**. The green spots are the **predicted RECIST diameters' endpoints by keypoint regression**.

this work, we also adopt the full HRNet for keypoint regression ($\text{HRNet}_{\text{regression}}$) by adding a global average pooling layer and a fully connected layer, and lesion segmentation and heatmap prediction ($\text{HRNet}_{\text{heatmap}}$) by adding a UNet-like decoder and skip connections. Table 2 lists the results of different settings. We observe that (1) the performance of all tasks are consistently improved after gradually introducing different components, demonstrating the effectiveness and contributions of each component for these tasks. (2) After simply adding a transformer encoder ($\text{HRNet}^* + \text{TranE}$), the performance gets a big improvement compared to $\text{HRNet}_{\text{heatmap}}$, meaning that the encoder can well capture the long-range dependencies of low-level features to enhance their representation ability. (3) After introducing the consistency losses, all tasks' results are improved, especially for keypoint regression when adding $\mathcal{L}_{\text{cons1}}$ and lesion segmentation when adding $\mathcal{L}_{\text{cons2}}$. Fig. 3 provides visual evidences to show the effectiveness of the proposed consistency losses, where the results of lesion segmentation and keypoint regression become closer to manual annotations and the predicted heatmaps become more concentrated and reliable after adding them.

Downstream Tasks. Two straightforward downstream clinic-relevant tasks can be conducted based on our results, which are 3D lesion segmentation and RECIST assessment in longitudinal studies. For 3D lesion segmentation, 200 lesions selected from above manually segmented 1,000 lesions are fully annotated with 3D masks following [4]. We run our system on these lesions slice-by-slice and stack their 2D segmentation results to form their 3D segmentations. This work achieves a mean Dice score of 85.6% remarkably surpassing the previous work [4] by 9.2%. For RECIST assessment in longitudinal studies, 480 lesion pairs from the test set of the DLS dataset [5]² are used for evaluation, which are excluded from our training data. The CT scans of each pair are collected at different time points from the same patient. For each lesion pair, we first get the lesions' RECIST diameters using our system. Then we compute its tumor response class, as

² <https://github.com/JimmyCai91/DLT>

complete response, partial response, progressive disease, or stable disease, of the manual RECIST and our automated RECIST assessments, based on RECIST version 1.1 [10]. On the 480 test lesion pairs, we achieve an accuracy of 91.7% for the tumor response classification, suggesting that our system can help clinicians to annotate sequential RECIST measurements with minimal human effort.

4 Conclusions

In this work, we propose a transformer-based network for lesion RECIST diameter prediction and segmentation. It leverages the long-range dependencies learning ability of transformers to enhance features. Besides, three tasks are performed and jointly trained, and their relationships are explicitly built by introducing two new consistency losses. All of these facilitate our system to produce accurate and robust results. Hence, it can serve as a useful tool for clinicians to precisely measure tumor sizes with minimal effort and may provide high positive clinical values on oncologic imaging analysis workflows.

References

1. Agarwal, V., Tang, Y., Xiao, J., Summers, R.M.: Weakly supervised lesion co-segmentation on CT scans. In: ISBI. pp. 203–206 (2020)
2. Agarwal, V., Tang, Y., Xiao, J., Summers, R.M.: Weakly-supervised lesion segmentation on ct scans using co-segmentation. In: Medical Imaging: Computer-Aided Diagnosis. vol. 11314, p. 113141J (2020)
3. Bretsch, M., Fränze, A., Merz, M., Hillengass, J., Semmler, W., Bendl, R., Bäuerle, T.: Assessing treatment response of osteolytic lesions by manual volumetry, automatic segmentation, and RECIST in experimental bone metastases. *Academic radiology* **21**(9), 1177–1184 (2014)
4. Cai, J., Tang, Y., Lu, L., Harrison, A.P., Yan, K., Xiao, J., Yang, L., Summers, R.M.: Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3D mask generation from 2D RECIST. In: MIC-CAI. pp. 396–404 (2018)
5. Cai, J., Tang, Y., Yan, K., Harrison, A.P., Xiao, J., Lin, G., Lu, L.: Deep lesion tracker: Monitoring lesions in 4d longitudinal imaging studies. In: CVPR. pp. 15159–15169 (2021)
6. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020)
8. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

10. 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)
11. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnU-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486* (2018)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
13. Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., Goh, R.S.M.: Medical image segmentation using squeeze-and-expansion transformers. In: *IJCAI*. pp. 807–815 (2021)
14. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: *NeurIPS*. pp. 8024–8035 (2019)
15. Rahman, M.A., Wang, Y.: Optimizing intersection-over-union in deep neural networks for image segmentation. In: *International Symposium on Visual Computing*. pp. 234–244 (2016)
16. Rothe, J.H., Grieser, C., Lehmkuhl, L., Schnapauff, D., Fernandez, C.P., Maurer, M.H., Mussler, A., Hamm, B., Denecke, T., Steffen, I.G.: Size determination and response assessment of liver metastases with computed tomography—comparison of RECIST and volumetric algorithms. *European Journal of Radiology* **82**(11), 1831–1839 (2013)
17. Tang, Y., Cai, J., Lu, L., Harrison, A.P., Yan, K., Xiao, J., Yang, L., Summers, R.M.: CT image enhancement using stacked generative adversarial networks and transfer learning for lesion segmentation improvement. In: *MLML*. pp. 46–54 (2018)
18. Tang, Y., Cai, J., Yan, K., Huang, L., Xie, G., Xiao, J., Lu, J., Lin, G., Lu, L.: Weakly-supervised universal lesion segmentation with regional level set loss. In: *MICCAI*. pp. 515–525 (2021)
19. Tang, Y., Harrison, A.P., Bagheri, M., Xiao, J., Summers, R.M.: Semi-automatic RECIST labeling on CT scans with cascaded convolutional neural networks. In: *MICCAI*. pp. 405–413 (2018)
20. Tang, Y., Yan, K., Cai, J., Huang, L., Xie, G., Xiao, J., Lu, J., Lin, G., Lu, L.: Lesion segmentation and RECIST diameter prediction via click-driven attention and dual-path connection. In: *MICCAI*. pp. 341–351 (2021)
21. Tang, Y., Yan, K., Xiao, J., Summers, R.M.: One click lesion RECIST measurement and segmentation on CT scans. In: *MICCAI*. pp. 573–583 (2020)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *NeurIPS* **30** (2017)
23. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* (2020)
24. 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* (3), 036501 (2018)
25. Yoon, S.H., Kim, K.W., Goo, J.M., Kim, D.W., Hahn, S.: Observer variability in recist-based tumour burden measurements: a meta-analysis. *European journal of cancer* **53**, 5–15 (2016)
26. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: *MICCAI*. pp. 14–24 (2021)