LOMIA-T: A Transformer-based LOngitudinal Medical Image Analysis framework for predicting treatment response of esophageal cancer

Yuchen Sun[1], Kunwei Li[2], Duanduan Chen[1], Yi Hu[3], and Shuaitong Zhang[1(✉)]

[1] School of Medical Technology, Beijing Institute of Technology, Beijing, P.R. China
zhangshuaitong@bit.edu.cn
[2] Department of Radiology, The Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai, Guangdong, P.R. China
[3] Department of Thoracic Surgery, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong, P.R. China

Abstract. Deep learning models based on medical images have made significant strides in predicting treatment outcomes. However, previous methods have primarily concentrated on single time-point images, neglecting the temporal dynamics and changes inherent in longitudinal medical images. Thus, we propose a Transformer-based longitudinal image analysis framework (LOMIA-T) to contrast and fuse latent representations from pre- and post-treatment medical images for predicting treatment response. Specifically, we first design a treatment response-based contrastive loss to enhance latent representation by discerning evolutionary processes across various disease stages. Then, we integrate latent representations from pre- and post-treatment CT images using a cross-attention mechanism. Considering the redundancy in the dual-branch output features induced by the cross-attention mechanism, we propose a clinically interpretable feature fusion strategy to predict treatment response. Experimentally, the proposed framework outperforms several state-of-the-art longitudinal image analysis methods on an in-house Esophageal Squamous Cell Carcinoma (ESCC) dataset, encompassing 170 pre- and post-treatment contrast-enhanced CT image pairs from ESCC patients underwent neoadjuvant chemoradiotherapy. Ablation experiments validate the efficacy of the proposed treatment response-based contrastive loss and feature fusion strategy. The codes will be made available at https://github.com/syc19074115/LOMIA-T.

Keywords: Longitudinal Medical Images · Treatment Response Prediction · Contrastive Loss · Feature Fusion · Esophageal Cancer.

1 Introduction

The analysis of longitudinal medical images is an important but challenging problem in the monitoring of tumor progression and the evaluation of treatment

response [10, 29]. Compared to single time-point medical images, longitudinal medical images offer more information related to treatment response [4]. For instance, neoadjuvant chemoradiotherapy (nCRT) followed by surgical resection is the preferred treatment for locally advanced esophageal squamous cell carcinoma (ESCC) [21]. For patients who achieved pathological complete response (pCR) after nCRT, a wait-and-see strategy is more suitable than surgical resection [7]. Therefore, preoperative prediction of pCR for individual ESCC patients is highly desirable, as it can assist clinicians in making treatment decisions.

Previous studies have demonstrated using deep learning-based or radiomics-based methods to analyze contrast-enhanced CT images can preoperatively predict pCR of ESCC [18]. Nevertheless, the majority of studies primarily concentrate on single time-point CT images (pre- or post-treatment), neglecting the temporal dynamics and alterations which can be elucidated through longitudinal contrast-enhanced CT images [12]. Several methods for predicting treatment response based on longitudinal medical images have been proposed, demonstrating that disease progression patterns represented by longitudinal data can improve the performance [11,15]. These works can be primarily categorized into deep feature contrast (DFC) based methods and deep feature fusion (DFF) based ones. Both of them usually establish two or more parallel neural networks to represent longitudinal images. Afterward, the DFC-based methods utilize a similarity function to quantify the differences among features from longitudinal images [5, 22], which are usually correlated to the treatment response. It is similar to the way clinicians evaluate tumor treatment response where they typically rely on visual comparison between pre- and post-treatment CT images. Thus, DFC-based methods are clinically interpretable, which is crucial for clinical decision-making. In contrast, the DFF-based methods usually fuse the features from longitudinal images and then associate the fused feature with treatment response. Compared with DFC-based methods, DFF-based methods for longitudinal images can provide a more comprehensive and accurate analysis of longitudinal changes [19,24]. Common feature fusion strategies include concatenation of features with all- or cross-attention mechanism [1], and their effectiveness have been validated in various tasks. For example, Tong et al. [23] developed a Dual-input Vision Transformer (DiT) model with the all-attention and concatenation strategy to fuse features from pre- and post-treatment medical images in breast cancer patients. However, self-attention leads to each output token encompassing features from every token in the longitudinal images. Concatenating all longitudinal tokens directly may introduce redundancy into the final feature pool, posing a potential impact on predictive performance. Disentangled representation learning is a method reducing the redundancy in fused feature pool with aligning longitudinal data in the time dimension [20] or latent representation dimension [28] to capture meaningful aspects. For instance, Yue et al. [26] proposed Multi-loss disentangled representation learning (MLDRL) to highlight differences and align commonalities among longitudinal images.

In this paper, we introduce LOMIA-T, a novel longitudinal image analysis framework that combines feature contrast and fusion techniques, through which,

we can effectively leverage longitudinal medical images for predicting treatment response. Specifically, we develop a clinically interpretable fusion strategy for longitudinal medical images and introduce a treatment response-based contrastive loss function aimed at capturing nuances in disease progression trajectories. Experimental results on an in-house ESCC dataset from two hospitals demonstrate that LOMIA-T outperforms other state-of-the-art longitudinal image analysis methods. Furthermore, LOMIA-T can be transferred for predicting treatment response in other diseases and can accommodate multiple time-point medical images with minor modifications.

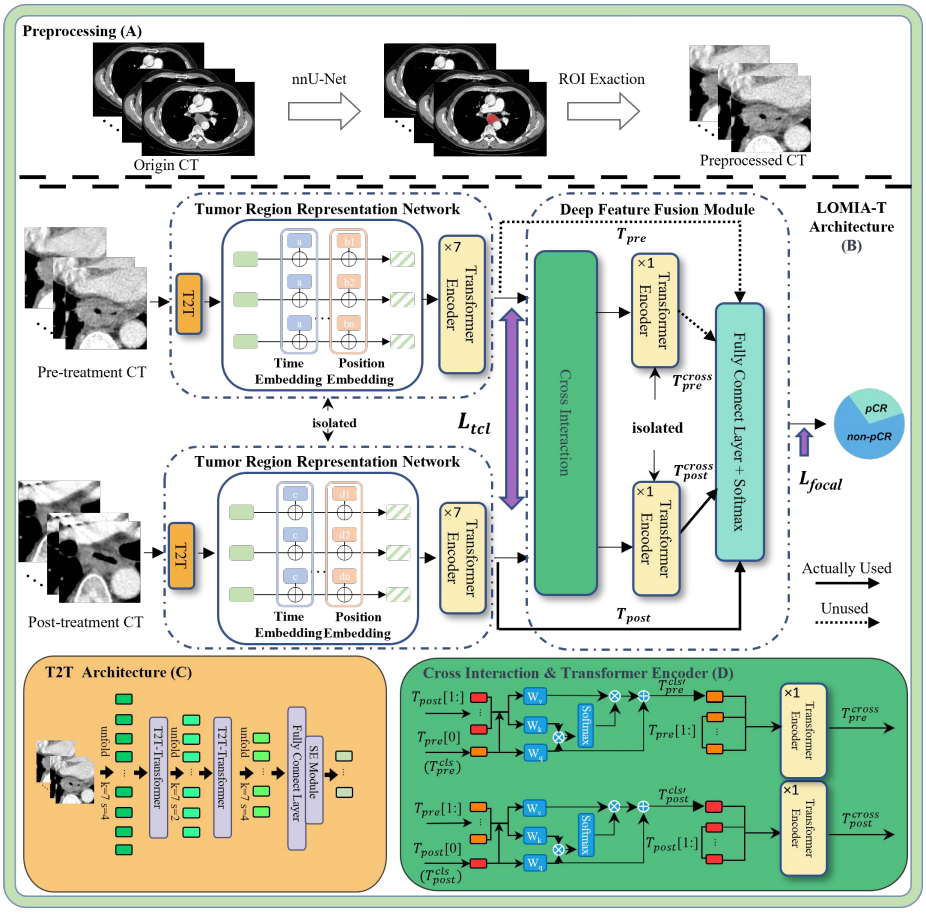

Fig. 1. Schematic of LOMIA-T. Solid lines represent T_{post} and T_{post}^{cross} which are used in LOMIA-T, while dashed lines represent T_{pre} and T_{pre}^{cross} which are not.(A) Preprocessing workflow; (B) Overall structure of LOMIA-T; (C) Schematic of the T2T module structure within the Tumor Region Representation Network; (D) Cross-attention in pre- and post-treatment branch.

2 Materials and Methods

2.1 CT Imaging and Preprocessing

This study was approved by the Ethics Committee of the Fifth Affiliated Hospital of Sun Yat-sen University (2021-K62-1) and was conducted in accordance with the Declaration of Helsinki. Informed consent was waived due to the observational design of this study and the de-identifed nature of the data. A total of 170 locally advanced ESCC patients who underwent nCRT are obtained from two hospitals, including 85 patients who achieved pCR, and longitudinal contrast-enhanced CT images (pre- and post-treatment) for these patients are available. We randomly divide these patients into a training and a test set at a ration of 4 : 1. We first use nnU-Net [9] to segment tumor region on the pre- and post-treatment contrast-enhanced CT images automatically. This method achieved a median Dice similarity coefficient (95 % confidence interval [CI]) value of 0.865 (0.851-0.875) on 580 eligible ESCC patients collected from two institutions. (Figure 1A) [27]. To guarantee the acquisition of the whole tumor region, we expand the three-dimensional bounding box of the segmented mask by 4 pixels on the cross-sectional slices and select 32 consecutive slices as the region of interest (ROI), fianlly resized the ROI into a fixed size of $32 \times 48 \times 48$.

2.2 Overall Network Architecture

Fig. 1B illustrates the network architecture of LOMIA-T. It mainly contains two subnetworks: tumor region representation network and deep features interaction network for pre- and post-treatment CT images. We propose a treatment response-based contrastive loss, enhancing the ability of the tumor region representation network to discern feature disparities indicative of treatment effects between pre- and post-treatment CT scans. Furthermore, a novel deep feature fusion method is introduced to improve prediction performance further.

Tumor region representation network. The tumor region representation network is constructed using a token-to-token Vision Transformer (T2T-ViT) module, and the representation networks for pre- and post-treatment contrast-enhanced CT images are structurally identical and mutually independent. T2T module is utilized to transform the input $X_{pre}, X_{post} \in \mathbb{R}^{D \times W \times H}$ (D = 32, W = 48, and H = 48 here) into tokens instead of the hard split used in the ViT [25], and it can retain textural and structural information. Each T2T module involves three steps: soft split (SS), T2T-Transformer (TT), and re-structurization (RS), with the step size in the soft split set to (4, 2, 2) and the same scale as the ROI while re-structurization. The output dimension in the T2T module is halved using a fully connected layer to reduce the complexity of the representation network and improve its generalization ability. Figure 1C illus-trates the T2T structure. The iterative process in T2T module can be formulated as:

$$I_{i+1} = SS\left(RS\left(TT\left(I_i\right)\right)\right), \ i = 1, 2, 3 \tag{1}$$

We apply a soft split at first to split X_{pre}, X_{post} to tokens: $I_1 = SS\left(X_i\right),\ i = pre, post$. The final iteration results in the output tokens $F_{pre}^{t2t},\ F_{post}^{t2t} \in \mathbb{R}^{l \times c}$, where c represents the number of channels and l signifies the output dimension.

In longitudinal images modelling, spatial and feature misalignments of the ROI on the longitudinal images always occur [16]. To mitigate spatial misalignment, two learnable matrices $PE_{pre},\ PE_{post} \in \mathbb{R}^{l \times c}$ are introduced as positional embedding, respectively. Positional embeddings describe the spatial relationships in images, while temporal embeddings indicate the temporal changes in images. Previous methods involve temporal embedding usually introduce a different temporal variable for different tokens [23]. We assume that tokens from the same time-point image share the same temporal embedding. Here, two learnable temporal embedding variables $TE_{pre},\ TE_{post} \in \mathbb{R}^1$ are also introduced for F_{pre}^{t2t} and F_{post}^{t2t}, respectively. After spatial and temporal embeddings, tokens from pre- and post-treatment CT images can be expressed as:

$$F_i = F_i^{t2t} \ + \ PE_i + TE_i,\ i = pre, post \tag{2}$$

Where $F_{pre},\ F_{post} \in \mathbb{R}^{l \times c}$.

We map $TE_{pre},\ TE_{post} \in \mathbb{R}^1$ to $\mathbb{R}^{1 \times c}$ using the Python-specific broadcasting operation and then add them together. Afterward, a seven-layer transformer encoder is employed to extract high-level imaging features $T_{pre},\ T_{post} \in \mathbb{R}^{l \times c}$.

Treatment response-based contrastive loss (TCL). Contrastive loss aims to reduce the distance between similar input pairs and increase that between dissimilar pairs in the feature space [3, 6]. For all ESCC patients, CT imaging features will change after nCRT. However, pCR group (with no tumor cells remained after nCRT) experiences a totally tumor regression, while non-pCR group experiences a smaller tumor regression or even tumor progression. In the task of predicting pCR of ESCC patients, the pre- and post-treatment CT images for patients who achieve pCR are considered as dissimilar pairs, and the pre- and post-treatment CT images for patients who does not achieve pCR are considered as similar pairs. Accordingly, we design a treatment respones-based contrastive loss (TCL) for predicting pCR:

$$\mathcal{L}_{tcl} = y\left(D_w\right)^2 + (1-y)\, max(m - D_w\ ,\ 0)^2 \tag{3}$$

Where y represents the patient's treatment response label (pCR = 0, non-pCR = 1), $D_w = \parallel T_{pre} - T_{post} \parallel_2$ denotes the Euclidean distance between features from pre- and post-treatment CT images. For pCR patients, we used the penalty only if D_w is smaller than m where m equals 0.5.

Deep Feature Fuison Module. The fusion module integrates tumor representations from pre- and post-treatment CT images using a cross-attention mechanism [1], as illustrated in Fig. 1D. More importantly, we consider the redundancy between tokens after the cross-attention mechanism and design skip

connections based on clinical prior knowledge to reuse low and high level semantic information for predicting pCR.

Specifically, for post-treatment CT branch, it first collects the patch tokens ($T_{pre}[1:]$) from the pre-treatment CT branch and concatenates its own CLS tokens ($T_{post}[0]$) to $T'^{post} \in \mathbb{R}^{l \times c}$. The mechanism then performs cross-attention between T_{post}^{cls} and T'^{post} with multiple heads (MCA), where CLS token is the only query as the information of patch tokens are fused into CLS token. The output $T_{post}^{cls\prime}$ with layer normalization and residual shortcut is defined as follows:

$$T_{post}^{cls'} = T_{post}^{cls} + MCA\left(LN\left(T'^{post}\right)\right) \tag{4}$$

Afterward, the CLS token and patch tokens from post-treatment CT branch are fed into a single layer transformer encoder including multi-head self-attention (MSA) mechanism and feed-forward network (FFN):

$$T_{post}^{cross} = \ FFN\left(MSA\left(T_{post}^{cls'} \parallel T_{post}\left[1:\right]\right)\right) \tag{5}$$

Where $T_{post}^{cross} \in \mathbb{R}^{l \times c}$.

Considering the redundancy between T_{pre}^{cross} and T_{post}^{cross}, their direct concatenation may be not the optimal feature fusion strategy. Previous study demonstrated that the combination of post-treatment CT imaging features and their interaction features with pre-treatment CT imaging features perform best for predicting pCR [14], which is consistent with the perception of clinicians. Inspired by the clinical prior knowledge and our previous finding, we concatenate T_{post}^{cross} and T_{post} with skip connection for predicting pCR. In this way, LOMIA-T can better utilize low and high level semantic information for predicting pCR. Specifically, the first token in $T_{post}^{cross}$and T_{post} is considered as the CLS token, respectively. These two CLS tokens are then processed through a fully connected layer with Softmax activation to create a binary classifier, enabling the computation of pCR classification results. In subsequent ablation studies, we also compare the performance of our fusion strategy with others.

Focal loss [17] is used as loss function to emphasize challenging samples here, denoted as $\mathcal{L}_{focal}$:

$$\mathcal{L}_{focal} = -\alpha\left(1-p_t\right)^{\gamma}\log\left(p_t\right),\ p_t = \begin{cases} p, & y=1 \\ 1-p, & y=0 \end{cases} \tag{6}$$

Where, $y = 1$ represents the non-pCR class, $y = 0$ represents the pCR class and $p \in [0, 1]$ is the probability for the class with label $y = 1$. Here, α balances the importance of pCR/non-pCR examples, and γ represents a tunable focusing hyperparameter. Additionally, α and γ are set to [1,1] and 2, respectively. The overall loss $\mathcal{L}$ is a weighted combination of $\mathcal{L}_{focal}$ and $\mathcal{L}_{ccl}$, and w_1 is set to 0.01 to ensure that these two losses are maintained at a comparable scale.

$$\mathcal{L} = w_1\mathcal{L}_{\ tcl} + (1-w_1)\mathcal{L}_{focal} \tag{7}$$

3 Results and Discussion

3.1 Experimental Settings and Evaluation

We train the LOMIA-T model using ten-fold cross-validation for 100 epochs (5 warm-up epochs) on A100 GPU with a batch size of 64. Other setup includes a cosine linear-rate scheduler with linear warm-up, an initial learning rate of 0.002, a weight decay of 0.05 and a DropKey [13] of 0.4. Moreover, we also apply the individual discrimination task to pre-train the tumor region representation net-work for pre- and post-treatment CT images, respectively [8]. Area under the curve (AUC), accuracy (ACC), sensitivity (SEN), and specicity (SPE) are used to evaluate the classification performance on the test set.

3.2 Comparison with Existing Methods

When comparing our proposed model LOMIA-T with three other deep learning methods for longitudinal images, LOMIA-T achieves superior performance in predicting pCR, as illustrated in Table 1. These methods include the feature contrast-based method DiT [23] and the feature contrast-based method Siamese-CNN [2] which implemented using their official codes, and the MLDRL based on the combination of feature contrast and fusion [26], with the experimental results sourced from the original study. Specifically, LOMIA-T yields an area under the curve (AUC) of 0.886, significantly outperforming DiT and Siamese-CNN ($p = 0.009$ and 0.006, respectively; DeLong test). It is important to note that parameters of DiT and Siamese-CNN are meticulously adjusted to optimize prediction performance using ten-fold cross-validation. Specially, Siamese-CNN adopts the framework of MoCo v2 [2], with a tumor progression task as pretext, which is associated with pCR labels. Additionally, to ensure comparability in model parameters with LOMIA-T, ResNet-18 is utilized as the backbone for Siamese-CNN. We train Siamese-CNN for 100 epochs, with the TCL as loss function. The optimal threshold m in the loss function is set to 1.4 using the same approach as that for LOMIA-T. All selected hyperparameters were optimized experimentally.

Table 1. Comparison of our model with existing methods for predicting pCR.

Method	AUC	ACC	SEN	SPE
DiT [15]	0.727	0.677	0.550	0.857
MLDRL [18]	0.866	0.810	0.684	**0.875**
Siamese-CNN [28]	0.529	0.588	0.750	0.357
LOMIA-T (Ours)	**0.886**	**0.824**	**0.800**	0.857

Table 2. Ablation studies on different strategies for expert classification and effectiveness of TCL loss function on longitudinal ESCC 3DCT dataset.

	Strategies	AUC	ACC	SEN	SPE
(a)	$\boldsymbol{T_{post} - T_{post}^{cross}}$ **(LOMIA-T)**	**0.886**	**0.824**	0.800	**0.857**
(b)	$\boldsymbol{T_{post} - T_{pre}^{cross}}$	0.879	0.765	0.750	0.786
(c)	$\boldsymbol{T_{pre} - T_{post}^{cross}}$	0.850	0.765	0.750	0.786
(d)	$\boldsymbol{T_{pre} - T_{pre}^{cross}}$	0.864	0.735	0.750	0.714
(e)	$\boldsymbol{T_{pre}^{cross} - T_{post}^{cross}}$	0.843	0.735	0.650	0.857
(f)	**w/o Deep Feature Fusion module**	0.711	0.735	**0.850**	0.571
(g)	**Only pre-treatment images**	0.546	0.500	0.350	0.714
(h)	**Only post-treatment images**	0.693	0.559	0.450	0.714
(i)	**w/o TCL**	0.821	0.677	0.600	0.786

3.3 Ablation Study

Several ablation experiments are also performed to verify the effectiveness of deep feature fuison strategy and treatment response-based contrastive loss, and the results are listed in Table 2. The training strategies and hyperparameters settings for the ablation studies are the same as LOMIA-T.

The effectiveness of deep feature fusion strategy. For predicting pCR, we employ a range of strategies to integrate features from pre- and post-treatment CT images, including the fusion of the features from the pre-/post- treatment after cross interaction $T_{pre}^{cross}/T_{post}^{cross}$ and features from pre-/post-treatment before cross interaction T_{pre}/T_{post}. Our experiments demonstrate that fusing features from T_{pre} or T_{post} with those from T_{pre}^{cross} or T_{post}^{cross} yield superior performance compared to the fusion of T_{pre}^{cross} and T_{post}^{cross}, the common strategy employed in previous studies [23]. This may be attributed to the observation that a higher degree of redundant information exists between tokens T_{pre}^{cross} and T_{post}^{cross} after cross-attention. When comparing the results presented in row (a) and (c), (b) and (d) in Table 2, we observe that imaging features extracted from post-treatment CT images exhibit greater predictive value than those from pre-treatment images. Moreover, in comparison to pCR prediction models based solely on pre-/post-treatment CT images, models integrating both pre- and post-treatment CT images consistently outperform them across all fusion strategies. These findings underscore the importance of longitudinal data in enhancing predictive accuracy for pCR.

The effectiveness of treatment response-based contrastive loss. We remove the TCL and employ only focal loss to train the LOMIA-T. Compared with the focal loss alone, the AUC and accuracy values are increased by 6.5% and 14.7%, respectively after adding the TCL. Using TCL forces the tumor representation network to extract pCR-associated imaging features from pre- and post-treatment CT images, thereby improving the pCR prediction performance.

4 Conclusion

In this study, we present a Transformer-based framework for longitudinal image analysis to predict treatment response of ESCC patients. We propose a clinically interpretable fusion strategy of longitudinal medical images and experimental results demonstrate its superiority over direct concatenation strategy. Moreover, the proposed treatment response-based contrastive loss can facilitate the model to capture the longitudinal changes of tumor on CT images, leading to a further prediction improvement. The LOMIA-T outperforms other state-of-the-art longitudinal image analysis methods on an in-house ESCC dataset from two hospitals and can be transferred for predicting treatment response in other diseases. In the future, we will further investigate the generalization of LOMIA-T on other diseases and the potential of LOMIA-T in predicting treatment response using multiple time-point medical images. Additionally, we will validate the effectiveness of LOMIA-T in larger scale longitudinal medical imaging datasets.

Acknowledgments. This study is supported by Beijing Natural Science Foundation (L232132) and National Nature Science Foundation of China (82102140).

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 357–366 (2021)
2. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
3. Chen, X., Fan, X., Chiu, B.: Interpretable deep biomarker for serial monitoring of carotid atherosclerosis based on three-dimensional ultrasound imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 295–305. Springer (2023)
4. Dadsetan, S., Arefan, D., Berg, W.A., Zuley, M.L., Sumkin, J.H., Wu, S.: Deep learning of longitudinal mammogram examinations for breast cancer risk prediction. Pattern recognition **132**, 108919 (2022)
5. Gu, J., Tong, T., He, C., Xu, M., Yang, X., Tian, J., Jiang, T., Wang, K.: Deep learning radiomics of ultrasonography can predict response to neoadjuvant chemotherapy in breast cancer at an early stage of treatment: a prospective study. European radiology pp. 1–11 (2022)
6. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006)
7. van Hagen, P., Wijnhoven, B., et al.: Recurrence pattern in patients with a pathologically complete response after neoadjuvant chemoradiotherapy and surgery for oesophageal cancer. Journal of British Surgery **100**(2), 267–273 (2013)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)

9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
10. Jin, C., Yu, H., et al.: Predicting treatment response from longitudinal images using multi-task deep learning. Nature communications **12**(1), 1851 (2021)
11. Konwer, A., Xu, X., Bae, J., Chen, C., Prasanna, P.: Temporal context matters: Enhancing single image prediction with disease progression representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18824–18835 (2022)
12. Lei, B., Yang, M., Yang, P., et al.: Deep and joint learning of longitudinal data for alzheimer's disease prediction. Pattern Recognition **102**, 107247 (2020)
13. Li, B., Hu, Y., Nie, X., Han, C., et al.: Dropkey for vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22700–22709 (2023)
14. Li, K., Zhang, S., Hu, Y., Cai, A., et al.: Radiomics nomogram with added nodal features improves treatment response prediction in locally advanced esophageal squamous cell carcinoma: A multicenter study. Annals of Surgical Oncology **30**(13), 8231–8243 (2023)
15. Li, X., Gao, H., Zhu, J., Huang, Y., et al.: 3d deep learning model for the pretreatment evaluation of treatment response in esophageal carcinoma: a prospective study (chictr2000039279). International Journal of Radiation Oncology* Biology* Physics **111**(4), 926–935 (2021)
16. Liang, W., Zhang, K., Cao, P., Zhao, P., Liu, X., et al.: Modeling alzheimers disease progression from multi-task and self-supervised learning perspective with brain networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 310–319. Springer (2023)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
18. Menon, N., Guidozzi, N., et al.: Performance of radiomics-based artificial intelligence systems in the diagnosis and prediction of treatment response and survival in esophageal cancer: a systematic review and meta-analysis of diagnostic accuracy. Diseases of the Esophagus **36**(6), doad034 (2023)
19. Rivail, A., Schmidt-Erfurth, U., et al.: Modeling disease progression in retinal octs with longitudinal self-supervised learning. In: Predictive Intelligence in Medicine: Second International Workshop, PRIME 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 2. pp. 44–52. Springer (2019)
20. Schön, J., Selvan, R., Nygård, L., Vogelius, I.R., Petersen, J.: Explicit temporal embedding in deep generative latent models for longitudinal medical image synthesis. arXiv preprint arXiv:2301.05465 (2023)
21. Shapiro, J., et al.: Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (cross): long-term results of a randomised controlled trial. The lancet oncology **16**(9), 1090–1098 (2015)
22. Szeskin, A., Rochman, S., Weiss, S., Lederman, R., Sosna, J., Joskowicz, L.: Liver lesion changes analysis in longitudinal cect scans by simultaneous deep learning voxel classification with simu-net. Medical Image Analysis **83**, 102675 (2023)
23. Tong, T., Li, D., Gu, J., et al.: Dual-input transformer: An end-to-end model for preoperative assessment of pathological complete response to neoadjuvant chemotherapy in breast cancer ultrasonography. IEEE Journal of Biomedical and Health Informatics **27**(1), 251–262 (2022)

24. Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., Mak, R.H., Aerts, H.J.: Deep learning predicts lung cancer treatment response from serial medical imaging. Clinical Cancer Research **25**(11), 3266–3275 (2019)
25. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., et al.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 558–567 (2021)
26. Yue, H., Liu, J., Li, J., Kuang, H., Lang, J., Cheng, J., Peng, L., Han, Y., Bai, H., Wang, Y., et al.: Mldrl: Multi-loss disentangled representation learning for predicting esophageal cancer response to neoadjuvant chemoradiotherapy using longitudinal ct images. Medical image analysis **79**, 102423 (2022)
27. Zhang, S., Li, K., et al.: Deep learning for automatic gross tumor volumes contouring in esophageal cancer based on contrast-enhanced ct images: a multi-institutional study. International Journal of Radiation Oncology* Biology* Physics (2024). https://doi.org/10.1016/j.ijrobp.2024.02.035
28. Zhao, Q., Liu, Z., Adeli, E., Pohl, K.M.: Longitudinal self-supervised learning. Medical image analysis **71**, 102051 (2021)
29. Zhu, H.B., Xu, D., Ye, M., et al.: Deep learning-assisted magnetic resonance imaging prediction of tumor response to chemotherapy in patients with colorectal liver metastases. International Journal of Cancer **148**(7), 1717–1730 (2021)